Kristiania University College

Master of Information Systems
Business Analytics

Social Media Analytics

Leveraging Twitter data to understand the dynamics of social media interactions on cryptocurrencies

Student number: 874043

A report submitted in partial fulfillment of the requirement for the degree of Master of Information Systems: Business Analytics

Kristiania University College
Prinsens Gate 7-9
0152 Oslo
Norway

Abstract

Rapid technological change in the last decades has led to the emergence of new platforms and fields such as cryptocurrencies and social media data. Cryptocurrencies are decentralized digital currencies that use blockchain technology to create a secure and decentralized environment. In the decade since the inception of social media, it has created revolutions and connected people with interests. Social media platforms such as Twitter allow users worldwide to share opinions, emotions, and news. Twitter is one of the most used social media platforms worldwide. The social media platform has millions of users where tweets are continuously shared every second. Therefore, tweets are useful when a large amount of data is generated to conduct a social media analysis. In addition, Twitter is broadly utilized by investors and financial analysts to gather valuable information. Several studies have shown that the content posted on Twitter can predict the movement of cryptocurrency prices. However, limited research has been conducted on the dynamics of Twitter interactions on cryptocurrencies among users. By leveraging 1724328 tweets, this research aims to understand the dynamics of social media users' interactions on cryptocurrencies. Essentially by shedding light on larger cryptocurrencies contrary to smaller. The findings reveal that Twitter users are more positive than negative about cryptocurrencies. The analysis also shows an existing relationship between events and the interaction of users, where cryptocurrency-related events shift the emotion, sentiment, and discussion topics of the users. The thesis contributes to demonstrating the effectiveness of the Social set analysis framework to analyze and visualize a massive amount of social media data and user-generated data created on social media platforms such as Twitter.

Keywords: Cryptocurrency, Text analytics, Social Media Analytics, Twitter, Social Set Analysis

Acknowledgments

I would sincerely like to express gratitude to Kristiania University College for providing me with high-quality education and academic achievements through the master's program in Information Systems – Business Analytics. I would also like to thank my thesis supervisor for insightful comments and contributions to the thesis.

Last but not least, I would like to give sincere thanks to family and friends for their encouragement and support throughout the process.

Total number of words: 21940

Table of Contents

1 Introduction

The emergence and evolution of technology have significantly impacted the world where humans discover and gain knowledge at a fast pace. Cryptocurrency is an area of interest that has attracted much attention in the modern age. The first cryptocurrency introduced was Bitcoin in 2008 by the acronym Satoshi Nakamoto (Kraaijeveld and De Smedt, 2020). Its decentralized nature and ability to work seamlessly across borders made it the first cryptocurrency to gain widespread acceptance. Bitcoin has provoked the release of many other cryptocurrencies based on blockchain technology. Because of the rapid growth of cryptocurrencies, the importance of social media posts and related news articles have also increased. The development of digital cryptocurrencies has been the most contentious and obscure innovation in the modern global economy (Derbentsev et al., 2019). Cryptocurrencies have also gained popularity for their massive return in a short time. In the field of finance, cryptocurrencies have introduced the concept of digital currencies, which are used to verify the transfer of funds and regulate the generation of units of value. Cryptocurrencies are becoming more widely known globally. Various governments and financial organizations have acknowledged their increasing popularity. Due to the emergence of new technologies, both academic and professional researchers have shown interest in understanding the behavior of these new assets. Because of the popularity of cryptocurrencies, a huge amount of data has been collected by users on Twitter and other social media platforms. This data can be used to identify trends and patterns in the market.

The rise of social media has revolutionized how information is disseminated on the internet. Through these platforms, users can now share their thoughts and views with a wide audience (Bartov et al., 2018). Social media gives the public room to express sentiments about current topics and events. Due to the rise of social media, Twitter has become a major platform for various types of communication. This includes financial news and advice, as well as marketing tools for cryptocurrencies. The platform contains a huge amount of emotional intelligence data of its users. Unlike other forms of traditional media, user-generated content is typically anonymous. Because of user-generated content characteristics, it is more continuous and nimbly conducted than other sources, making it a more accessible information source (Bartov et al., 2018).

The thesis aims to study the online conversations surrounding cryptocurrencies by employing a data-driven approach to social media. Using 1724328 tweets and employing the theoretical lens of social set analysis, this study attempts to understand the dynamics of social media interactions[1] on cryptocurrencies. This study also aims to study the interactions on and between larger and smaller[2] cryptocurrencies. This thesis will attempt to answer the following research questions:

- What are the different social media user interactions around cryptocurrencies?
- How does the user interaction vary between larger coins and smaller coins?
- How are these user interactions affected by important events?

1.1 Delimitations of the Thesis

The research involves extracting Twitter data by utilizing Snscrape i.e., a Python library for accessing Twitter data without accessing APIs to collect and store tweets that mentions Bitcoin, Ethereum, Uniswap and Dogecoin. Therefore, the findings of this thesis are limited by data extracted and probably may not be generalized[3] to other social media platforms like Facebook, Instagram, LinkedIn and Redditt. User-generated content typically consists of text, images and videos (Ghani et al., 2019). The user-generated content in the thesis is limited to tweets that only include text data. Though Twitter is available in 33 languages, the data extraction was based on the English-language, where only Twitter posts in English were collected (Twitter.com). The findings related to smaller and larger coins are limited to the four chosen cryptocurrencies for this study, which are Bitcoin, Ethereum, Dogecoin, and Uniswap.

---

[1] The user interactions refer to the users' sentiments, emotions, and discussion topics in the cryptocurrencies. This study also employs sentiment and emotion analysis, which serves as a backdrop for further analysis of users and events.

[2] Smaller coins refer to cryptocurrencies with a lower market cap, and larger coins refer to cryptocurrencies with a higher market cap. The market cap of a particular cryptocurrency is a tool used to measure its overall potential and compare its value with other cryptocurrencies (Gomez, 2022).

[3] The audience using twitter and have conversations here will be different as compared to Facebook or LinkedIn.

1.2 Thesis Disposition

The structure of the thesis is like that of a traditional master thesis. The structure involves nine parts (counting introduction) – which includes a theoretical foundation, literature review, methodology, data, analysis, findings and discussion, limitation, and conclusion.

Chapter 2 – Theoretical foundation: This chapter describes concepts related to understanding social media analytics and the scope of cryptocurrencies.

Chapter 3 – Literature review: This chapter identifies and present literature on social media data and cryptocurrency and describes related work and previous research.

Chapter 4 – Methodology: This chapter present the research methodology and analytical approach for the thesis.

Chapter 5 – Data: This chapter describes the data extraction and collection for the thesis, as well as data pre-processing.

Chapter 6 – Analysis and findings: In the analysis section, the analyses will be conducted and presented based on the methodological approaches.

Chapter 7 – Discussion: In this chapter, the findings in the analyses will be presented and discussed and aim to answer the research questions.

Chapter 8 – Limitations: This chapter will mention the research's limitations.

Chapter 9 – Conclusion and implications: The last and final chapter will present the conclusion of the research, implications for practice and the research's contributions.

2 Theoretical Foundations

This section reviews the theoretical foundations of social media analytics, as well as describes concepts related to understanding the scope of cryptocurrencies. Therefore, the theoretical background will be categorized into two parts: Social media analytics and cryptocurrency.

2.1 Social Media Analytics

According to Google Trends data, the term social media analytics has been around since 2008. It has continuously increased since that time. Due to the increasing popularity of social media, Social media analytics is gaining prominence among various business communities and researchers. Social media analytics (SMA) is the method and approach of collecting data from social media sites and blogs to analyze and evaluate the data to make decisions or conduct research (Sponder and Khan, 2018, p. 170). It is the process of using the power of social media data to extract valuable insight. Social media analytics is also a science that systematically analyzes and extracts data from different social media platforms. While it provides valuable insights, it is also an art that requires deep knowledge of various business areas to implement effective strategies (Sponder and Khan, 2018, p. 170). The complexity of social media analytics involves the combination of advanced tools and technology and the possibility of analyzing and interpreting the data.

Social media analytics consists of layers such as social media text analytics. It is a type of social media data that focuses on extracting and analyzing the data from various social media platforms. It is commonly used to identify emerging trends and topics through insights into textual elements such as comments, tweets, and posts (Sponder and Khan, 2018, p. 173). Another layer is social media platforms such as Twitter, Facebook, and LinkedIn. Social media analytics is most commonly used to measure brand loyalty, increase data traffic to media, forecast, business intelligence, sentiment analysis, and decision-making (Sponder and Khan, 2018, p. 174). Social media analytics can be divided into descriptive, predictive, and prescriptive analytics (Sponder and Khan, 2018, p.175). Descriptive analytics analyzes social media text to identify user sentiments and emerging trends. It gathers and describes social media data in contexts such as visualizations and reports and is further based on action analytics and some aspects of text analytics (Sponder and Khan, 2018, p. 177). Predictive analytics analyzes large amounts of accumulated social media data to predict future patterns by analyzing intentions expressed in social media (Sponder and Khan, 2018, p. 178). Prescriptive analytics gives indications on the best option to act upon.

The cycle of Social media analytics is a six-step process that involves the science and art of mining insight from data generated by social media (Sponder and Khan, 2018, p. 176). The identification stage of Social media analytics is where the goal is to find the correct information to analyze. The vast amount of data available over social media platforms makes it challenging to analyze. Framing the right questions and approaches in terms of the data is essential to gaining valuable insight (Sponder and Khan, 2018, p. 176). As well as the channel the data is generated through, for example, Twitter data.

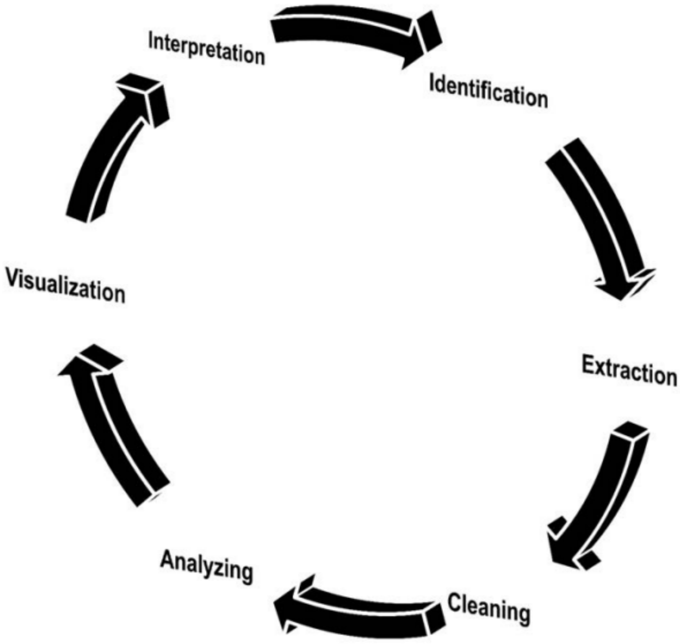
The second step in the cycle is extraction. This includes determining the type of data you want to extract and the platform tools to handle it. For instance, text data can be extracted using large-scale automated extraction through an application programming interface (API), through programming, or other automated processes (Sponder and Khan, 2018, p. 176). The extraction of social media data can follow privacy and ethical issues. Therefore, data extraction techniques should not violate a user's privacy and should be handled carefully. Social media platforms should clearly state their ownership of activities associated with data collection and processing (Sponder and Khan, 2018, p. 177).

The third step in the cycle is cleaning unnecessary data from the extracted data (Sponder and Khan, 2018, p. 178). Some data may need to be cleaned, while other data can be used directly for analysis. In the cases of text analytics, the cleaning may involve coding, clustering, and filtering using natural language processing.

The fourth step in the cycle includes analyzing the clean data to gain insight. Depending on the complexity of the query and the algorithm used, the approach may vary. Most of the time, the query comes across as a step-by-step process of analyzing a massive amount of data (Sponder and Khan, 2018, p. 178).

The fifth step is visualization. Visual analytics is essential for interpreting the analyzed data, and it is a factor in interactive decision-making (Sponder and Khan, 2018, p. 178). An effective tool is commonly used to visualize complex and big datasets, such as social media data. It can reveal hidden patterns and relationships within the data. Different forms of visualization may include multidimensional charts, graphs, tables, heat maps, plots, computer simulations, etc. (Sponder and Khan, 2018, p. 178).

The sixth and final step involves interpretation; this includes human judgment to interpret insight and knowledge from the visualized data (Sponder and Khan, 2018, p. 178). Two approaches often used are producing understandable analytical results and improving analytics to gain insight. The purpose is to make the data appear as if presented in a form that a person will use (Sponder and Khan, 2018, p. 178).



Figur 1. The Social media analytics cycle (Sponder and Khan, 2018, p. 176).

2.1.1 Social Media

The term social media refers to computer-based platforms that enable users to connect and communicate with each other. The information is distributed through virtual networks and communities. It allows users to store and share information such as photos, videos, and documents (Dollarhide, 2021). Social media is a two-way street that enables users to communicate. Its increasing popularity has created an environment where people can freely discuss and interact with each other (Ganis and Kohirkar, 2015, p. 20).

The use of social media as a platform to share information, knowledge, opinions, and emotions has increased rapidly in the last decade. Social media has exploded as a form of online discourse, and it allows people to create content, share it, and network at a massive rate. Social media's rapid emergence and evolution have become an integral part of society. The immense amount of data generated by social media platforms present an opportunity to develop models that predict future trends and provide targeted solutions (Asur and Huberman, 2010).

Social media platforms have become the primary source of information about cryptocurrencies, and Twitter and Reddit are amongst the biggest cryptocurrency-related forums. In such microblogging sites, it is common to reflect and share opinions. The messages on these sites are often concise and contain only a few words. The sentiment of the posts can be negative, positive, or neutral. This can be easily detected through an automated message detection system and is common for firms in relation to customer feedback on products and services. Social media influences public opinion, and mass media communication theories consider how media outlets can influence public opinion and how they can shape it. People's opinions and perceptions also influence the volatility of the cryptocurrency market.

Kraaijeveld and De Smedt (2020) state that due to the rise of social media platforms, investors are more likely to rely on them for accurate information about the cryptocurrency market. Cryptocurrency volatility is influenced by messages and posts on social media platforms (Kraaijeveld and De Smedt, 2020). Abraham et al. (2018) observed that the rise of social media and the increasing number of people using it impact the prices of cryptocurrencies. The more people communicate through social media, the more the impact on the consumer’s decisions on buying and selling increases. Social media posts and messages can easily influence the way people think and how they see the world. In this way, people may get motivated to take bigger risks without realizing that the market volatility could affect them (Abraham et al., 2018).

2.1.2 Social Media as Big Data

Social media data is high in volume, velocity, and diversity. The generated data is massive; however, capturing and analyzing all of it may be challenging. Big data can be characterized as a type of data with many characteristics. It refers to a very large and complex type of data, which makes it difficult or impossible to process through traditional methods (Ganis and Kohirkar, 2015, p.66). According to Forbes, the first use of the term big data was made in 1997 by scientists from NASA. They talked about the challenges they faced in analyzing and visualizing their massive datasets (Ganis and Kohirkar, 2015, p.65).

Big data has been around for quite some time, but it has only recently become more prevalent due to the rise of social media. With the help of this type of data, businesses and people can now analyze and visualize the various trends on the platform. Social media platforms are both communication tools and platforms for connectivity. They allow users to exchange different types of information. Some of these include a short text (Twitter), friendship networks (Facebook), and instant pictures (Instagram) (Elmer et al., 2015, p.4). The communication between people through social media generates social media data. Social media as big data are unstructured or semi-structured and contain comments, tweets, hyperlinks, images, emoticons, and videos. Therefore, social media data needs cleaning and transformation (Sponder and Khan, 2018, p. 172).

2.1.3 Twitter

Twitter has become a major platform for financial news and advice and a marketing tool for cryptocurrencies because of the public discussion on the social media platform. With the rise of social media, people have been able to express their emotions and concerns through various platforms. In particular, Twitter has gained popularity due to its 280 character status updates. Twitter is a social media platform that enables users to post and interact with "tweet" messages. Its users can also "like" and "retweet" the messages they see on the platform. Hashtags are also used to identify a tweet. They are followed by a string of characters. This feature is also used to collect data about a tweet. Not only do Twitter users regularly receive live updates about cryptocurrencies, but they also have the emotional intelligence to make informed decisions. The social media platform has 206 million users as of the second quarter of 2021 (Twitter).

Researchers have discovered the various reasons people use Twitter. These include finding common ground and connections and a channel to broadcast information (Zhao and Rosson, 2009). The behavioral economics in Abraham et al. (2018) states that sentiment and emotions can affect individual behavior and decision-making. Abraham et al. (2018) utilized Twitter to study tweet volumes and sentiments regarding price directions. Twitter has a large amount of data that contain the emotional intelligence of cryptocurrency users and investors. Twitter is a huge data source that provides information about any given topic. Therefore, the researchers argue that Twitter is a great place to collect text data, especially regarding cryptocurrencies, to explore the relationships between the sentiment in the text and prices. In addition.

2.2 Cryptocurrency

2.2.1 Background and concepts of cryptocurrency

Virtual currencies have made huge developmental leaps in the last couple of years. Earlier forms of these electronic currencies were very innovative in their ability to transfer large amounts of money at fast speeds (Bohr and Bashir, 2014). Bitcoin was first introduced in 2009 by Satoshi Nakamoto and is a decentralized digital currency that serves as a medium of exchange and a store of value. Its creation was followed by other virtual currencies. Cryptocurrency is a decentralized virtual currency that is based on blockchain technology. Cryptocurrency is a disruptive, decentralized, and cryptographic technology that enables the digitalization of trust (Kraaijeveld and De Smedt, 2020). It is also a source of exchange that uses cryptography to secure financial transactions (Naimy et al., 2021). Unlike fiat money, cryptocurrencies are not backed by any central authority and are not subject to regulation. This characteristic makes them unbanked and thus immune to the central banking system's interference. Therefore it provides advantages over traditional payment methods, such as speed, high liquidity, lower transaction costs, and anonymity (Naimy et al., 2021).

In 2008, a new decentralized digital currency called Bitcoin was launched by an anonymous author who used Satoshi Nakamoto as a name. Its launch coincided with the publication of Nakamoto's revolutionary white paper, which provided a secure and robust framework for establishing a digital currency. This formed the basis of blockchain technology (Kraaijeveld and De Smedt, 2020). Years later, other cryptocurrencies, often referred to as altcoins, such as Ethereum and Litecoin were developed. Over 5000 altcoins have been released, altcoins are cryptocurrencies other than Bitcoin. The total market capitalization of cryptocurrencies is $258 billion, Bitcoin alone has a market capitalization of $179 billion

Due to the rapid growth of their value, cryptocurrencies have been the subject of various academic and media interest. Their increasing popularity has prompted many central banks to explore their potential as a new type of financial asset (Aslan and Sensoy, 2019). Cryptocurrency is a general phenomenon characterized by volatility and abrupt price changes. The lack of regulation on the global level has created an environment where risk-taking is common. The world's two largest cryptocurrencies in the form of market capitalization are Bitcoin followed by Ethereum. Bitcoin is considered the first decentralized digital currency with direct transactions between users without an intermediary.

2.2.2 Blockchain

Blockchain is a distributed ledger technology that enables digital assets to be unalterable and transparent (Bohr and Bashir, 2014). Its decentralized nature allows people to store and record transactions. Cryptocurrencies are powered by blockchain technology. A cryptocurrency's blockchain is a master public ledger that records all transactions and activities in the currency. A blockchain is a record of all transactions that have occurred on it since it was created, and its finite length increases over time (Bohr and Bashir, 2014).

The critical difference between a typical database and a blockchain is the data structure. A blockchain collects data together in groups, known as blocks. When filled, the blocks are linked to a previously filled block, and new information is added to the chain (Hayes, 2022a). A database typically stores its data in rows and columns. On the other hand, a blockchain structures its data into blocks that are strung together. The data structure makes an irreversible timeline of data when implemented in a decentralized nature. When a block is filled it is eternal and becomes a part of this timeline. Every block in the chain is given an exact timestamp when added to the chain (Hayes, 2022a). The success of cryptocurrencies is primarily linked to their ability to reduce transaction costs. Some researchers believe that cryptocurrencies could provide users with more autonomy and lower costs (Mnif et al., 2021).

2.2.3 Decentralized Finance (DeFi)

DeFi stands for decentralized finance and is a system where financial products are available on a public decentralized blockchain network (Sharma, 2021). DeFi is a term that describes the various financial applications in cryptocurrency and blockchain which aims to disrupt financial intermediaries. DeFi is open for everyone to use and is not dependent on a public decentralized blockchain network. Unlike a bank account or brokerage account, DeFi-compatible software can be used to transact without requiring a government-issued ID or proof of address (Sharma, 2021). Most DeFi applications are built on Ethereum technology, such as Ethereum's platform for smart contracts. It is governed by smart contracts, which are digital self-executing and can be used to carry out transactions (Eaton). In this way, intermediaries are removed. DeFi tokens in 2021 can be used for many financial services, such as mortgages, loans, asset trading, and contracts, without needing an intermediate platform. In this paper, the focus in regards to DeFi coins will be on Uniswap.

2.2.4 Meme coins

Meme coins are digital tokens inspired by popular memes and jokes about cryptocurrencies. Meme coins are associated with a theme. They usually gain popularity in a short time and are often gain hyped by notable influencers on social media (Frankenfield, 2021). Meme coins are Altcoins, which are cryptocurrencies other than Bitcoin. They share similar characteristics with Bitcoin but stand out. Meme coins are crypto assets and are a category of cryptocurrencies and tokens that occurred due to a joke, generally derived from the internet (Thomas, 2021). Tesla CEO Elon Musk is known to frequently post cryptic tweets about the meme coin Dogecoin, which often moves their prices and causes fluctuations. In December 2021, Elon Musk tweeted that Tesla merchandise would be buyable with Dogecoin. The announcement generated a price increase of 20% (Browne, 2021). The meteoric rise of these cryptocurrencies was widely attributed to pure speculation (Frankenfield, 2021). In this thesis, the focus in regards to Meme coins will be on Dogecoin.

2.2.5 Selected coins

The selected coins for the research are Bitcoin, Ethereum, Dogecoin, and Uniswap. In terms of market capitalization, Bitcoin and Ethereum rank the number one and two globally. Dogecoin and Uniswap are considered smaller coins in market cap. The market cap of a cryptocurrency is the total value of its circulating supply. It is similar to that of the stock market's free-float capitalization.

Bitcoin

Bitcoin is considered the world's first blockchain technology and was founded in 2009 (Frankenfield, 2022). Since then, Bitcoin has become the benchmark for other cryptocurrencies. Bitcoin was created by the pseudonymous Satoshi Nakamoto. Bitcoin was founded on the idea of a peer-to-peer cash system and was developed using a decentralized network that's based on cryptography (Frankenfield, 2022). Bitcoin is created, distributed, traded, and stored on blockchain technology. Bitcoin has a market cap of approximately 550 billion USD per May 2022 (CoinMarketCap, 2022).

Ethereum

Ethereum is the world's second-largest cryptocurrency. Its main purpose is to provide a secure and transaction-oriented environment. Ethereum is built on Bitcoin's innovation, but with some significant differences (Ethereum.org). Ethereum is also a programmable digital money platform. It can be used to purchase digital goods and services without intermediaries (Ethereum.org). Ethereum is also a marketplace where people can buy and sell digital goods and services. Ethereum's technology is an expansion of the concept behind Bitcoin. As well as being a currency, it is also a global decentralized technology network. Ethereum has a market cap of approximately 235 billion USD per May 2022 (CoinMarketCap, 2022).

Dogecoin

Dogecoin is the biggest original meme coin at the moment. Launched in 2013, Dogecoin is an open-source cryptocurrency. Dogecoin is a cryptocurrency that uses blockchain technology and is a distributed digital ledger that enables secure transactions (Rodeck, 2021). Dogecoin can be used for making purchases and payments. Dogecoin reached an all-time high of $0.74 in May 2021 (Thomas, 2021). The creation of Dogecoin was a joke or a parody of bitcoin, but it quickly became a popular commodity. Its participation in the cryptocurrency bubble sent its values up significantly. Despite losing much of its value, it still has a following generated from posts on social media platforms such as Twitter and Reddit (Frankenfield, 2021). Dogecoin has a market cap of approximately 11 billion USD per May 2022 (CoinMarketCap, 2022).

Uniswap

Uniswap is a leading decentralized crypto exchange that uses the Ethereum blockchain (Leech, 2021). Uniswap was created in 2018 by Hayden Adams, a former mechanical engineer (Leech, 2021). Uniswap is a decentralized exchange that enables people to trade in $1 billion or more in daily crypto. Its market value is $12 billion (Leech, 2021). In traditional exchange, the price of an asset is regulated between the bid and ask. On DeFi coins such as Uniswap, prices are also based on market makers and their understanding of supply and demand (Angeris and Chitra, 2020). Uniswap has a market cap of approximately 3 billion USD per May 2022 (CoinMarketCap, 2022).

3 Literature review

This section will identify and appraise literature on social media data and cryptocurrency. First, the literature review search strategy will be presented, followed by a presentation of related work and previous research. Lastly, the purpose and objective of the research will be accounted for, as well as the knowledge gap and contribution.

3.1 Literature review search strategy

A systematic search was carried out to critically review the literature for the concepts and analysis in the thesis. The search process was carried out according to conditions set in the early stages. The goal of the process was to narrow down the list of possible sources of information to only be left with those that are relevant to the research.

Rousseau et al. (2008) suggested a four-step procedure when conducting a systematic literature review. This approach aims to provide a comprehensive and transparent analysis by identifying central literature by analyzing and interpreting the chosen studies (Rousseau et al., 2008). This method will be implemented as the literature review search strategy. The first step is to formulate research questions, which are further debated and revised until appropriate for the intended objectives (Rousseau et al., 2008). Second, comprehensive identification of relevant research is conducted by identifying relevant and high-quality literature from scientific search databases (Rousseau et al., 2008). Third, the research questions are used as the starting point for analyzing and selecting articles (Rousseau et al., 2008). Fourth, a synthesis of the findings is performed (Rousseau et al., 2008).

The literature search mainly focused on utilizing databases and considered journal articles and conference publications from three databases related to the field of computer science. The literature search was conducted by gathering secondary data from global leading academic journals and publications, focusing on peer-reviewed articles. Consequently, articles and publications were first and foremost gathered from the database of IEEE (Institute of Electrical and Electronics Engineers), ACM (Association for Computing Machinery) and Elviser/Science Direct, specifically "International Journal of Information Management Data Insights." In addition, the research database at Kristiania University College’s library, “Oria” was utilized. The research process was narrowed down by using terms and keywords to find articles that are directly connected to my field of research. Table 1 summarizes the search terms and keywords used for the systematic literature review. The search included identifying the most relevant articles that analyzed the social media data from 2016 to 2022. The field of social media analytics is a fast-moving and progressing field. For this reason the relatively short timeframe was chosen. The search process consists of a comprehensive and extensive systematic search to retrieve relevant literature.

Search terms and keywords	Databases	Fields
"Social media analytics" “Social media analysis” AND “cryptocurrency” “Social media data” AND “cryptocurrency” "“Big data analysis” AND “cryptocurrency”	IEEE ACM Elviser/Science Direct Oria	Title, Abstract, Keywords, Year

Table 1. Keywords and databases

Papers with relevance to the search were categorized according to:

- Relevance to the field of social media analytics
- Research related to cryptocurrency

The literature extracted from databases was then categorized according to its relevance. This method utilizes a ranking system that considers the number of search terms in each article. The relevance of an article was assessed by thoroughly examining titles, abstracts, keywords and year of the paper. Additionally, by using filters in the database the search was optimized. The additional strategy included searching in Google Scholar, to find literature in various academic disciplines. However, it was a point to sort by the number of citations because this shows the publication's reliability.

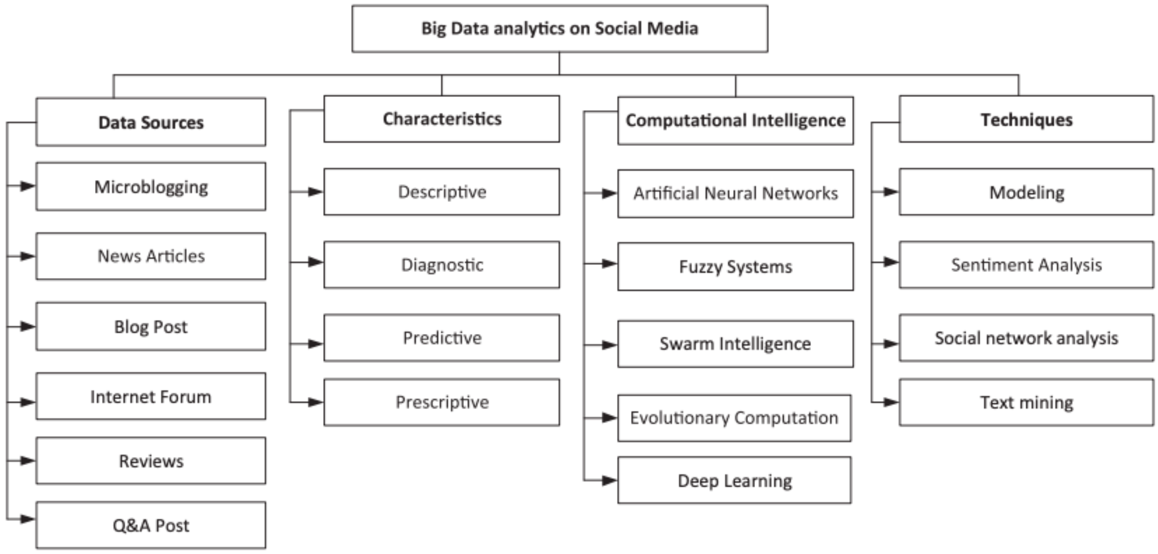
3.2 Related work

In this section, a brief review of related work will be presented using big data methods, machine learning and descriptive analysis of social media data. The state of the art related to the recent stages of social media analytics will also be presented.

3.2.1 Big data methods on social media data

The popularity of cryptocurrency has led to a huge amount of data generated from social media platforms such as Twitter. Big data methods can help identify trends and patterns in activities in the cryptocurrency market. The massive amount of data collected by users through these platforms has been the result of their daily activities and background details.

Researchers have developed a conceptual model based on the characteristics of analytics on social media. Figure 2 summarizes the classification of methods in big data analytics on social media. This conceptual model reflects the categorization of big data analytics based on previous research about social media data such as Kim and Hastak (2018), Stieglitz et al. (2018), and Ghani et al. (2019). The field of social media analytics has the objective of combining, extending, and adapting various approaches for analyzing data generated in social media.



Figur 2. Classification of Big Data Analytics on social media (Adopted from Ghani et al., 2019)

According to Zachlod et al. (2022), the majority of social media data stems from Twitter, while sentiment and content analysis are the current prevailing methods. Further, Ghani et al. (2019) state that 46% of current studies on social media data are analyzed on Twitter data rather than other social media sources. A total of 17% typically perform analysis on reviews, as well as views and opinions collected by users on the platform. Most studies on social media data also focus on microblogs as their primary source of data. They typically perform sentiment analysis, classification, and text analysis. Recent work on big data methods on social media is related to natural language programming, sentiment analysis, social network analysis, and news analytics (Ghani et al., 2019).

One of the most popular big data techniques used in social media analysis is text mining, which is a type of information extraction that can be performed from various types of unstructured content. Several studies on cryptocurrencies have been presented that analyze various aspects of technology, security, privacy, and applications. Although there isn't a text mining analysis precisely on profile analysis of cryptocurrencies, various studies focus on price predictions (Narman and Uulu, 2020).

Matilda (2017) stated that social media platforms has created an immense amount of data that can unlock the hidden insights of human behavior. Further, the author argues that big data analytics is a necessary component of any organization's strategy to analyze and improve its operations. For this reason, new tools and methods is increasing.

3.2.2 Machine learning approach

Studies have shown that most of the techniques used in analyzing the data collected by social media users rely on machine learning (Ghani et al., 2019). The machine learning approach on social media regarding cryptocurrency has essentially been used for price prediction. For instance, Mohapatra et al. (2019) proposed a distributed architectural design that combines machine learning and a lexicon approach to analyze the sentiment on Twitter to predict cryptocurrency price movements. The researchers in Munim et al. (2019) proposed testing ARIMA and neural network auto-regression (NNAR) models capable of forecasting the daily price movement in Bitcoin to improve the literature on cryptocurrencies. Wong (2021) utilized Twitter data and Natural Language Processing to study the influence of social media on Bitcoin prices. The study was conducted by using LSTM and Nave Bayes model to analyze the sentiment of tweets to predict the price signal of Bitcoin. Recently, Barradas et al. (2022) conducted a research based on k-means clustering that proposed architecture for processing cryptocurrency and social media data. The finding illustrated that activity on Twitter correlates with the action in the cryptocurrency market. The study also found that 14% of the data related to extraordinary behaviors are related to cryptocurrencies.

3.2.3 Descriptive analysis

Descriptive analysis in social media analytics uses historical data for analysis and visualization to provide valuable information. The method quantifies, identifies, and categorizes connections in the data (Ghani et al., 2019). One of the main aspects of social media analytics is the perspective of the users involved in creating the content. This perspective allows researchers to explore the various roles that social media users play in creating and disseminating information (Stieglitz et al., 2018).

Descriptive analysis has been cited by various researchers regarding social media analytics on cryptocurrency. Blau et al. (2021) provide a descriptive time-series analysis of inflation and Bitcoin, where the researchers are investigating the relationship between the time-series relation between the expectations of Bitcoin and the forward inflation rate. The study uses an autoregressive vector process. Khan and Hakami (2021) conducted a descriptive analysis (qualitative research) on cryptocurrency-related to usability perspective versus volatility threat. Rodrigues (2019) compares and contrasts the quantitative and qualitative methods used in social media data exploration. The research was conducted by a topic modeling approach for the descriptive analysis of extensive unstructured Twitter data.

3.2.4 Current state-of-the-art

Vatrapu et al. (2016) presented the framework of Social set analysis, which combines social data's theoretical and practical aspects. It also includes an analytical framework that combines big social data sets with societal and organizational data. Social set analysis (SSA) is based on the philosophical principles that stem from ecological psychology, micro-sociology, associational sociology, as well as mathematics in terms of set theory (Vatrapu et al., 2016).

The concept of SSA is a framework that combines the theoretical foundations of social science and computational social science (Vatrapu et al., 2016). It also provides an analytical framework for analyzing large social data sets. Twitter conversations create networks with recognizable shapes. The people who are driving the conversation can differ depending on the topic. The social connections on Twitter are composed of users and their relations with other users through text. According to Vatrapu et al. (2016), the basic premise of SSA is that actor A has an association with an entity E, which can be an artifact or an actor. Social set analysis is a framework for analyzing the interactions and builds on set theory. Vatrapu et al. (2016) argues that there is a lack in deep academic knowledge of the most dominant action that users perform on social media platforms every day.

Flesch et al. (2015) argue that the current state-of-the-art in big social data analytics is limited due to the focus on graph theoretical approaches such as Social Network Analysis (SNA). Current analytical approaches in social science can be divided into four main categories: text analysis, social network analysis, complex systems science, and social simulations (Vatrapu et al., 2016). Vatrapu et al. (2016) state a lack of framework for analyzing social media interactions that address the various units of analysis typically associated with these platforms. Therefore, the paper proposes a new approach called Social Set Analysis, which aims to provide a framework for analyzing and understanding these interactions. The study also concluded that SSA covers the scope of prescriptive, visual, and descriptive analytics.

3.2.5 Theory of Social data

The theory of social data in this thesis is based on the theory from Vatrapu (2010) where the author discuss sociotechnical interactions. The concept of sociotechnical interactions is conceptualized as a combination of the perception and appropriation of various socio-technical structures and technological functions (Vatrapu, 2010) These include the socio-technical affordances and the technological intersubjectivity (Vatrapu, 2010). The concept of technological intersubjectivity refers to the interactions between two or more users. Twitter as a social media platform includes individuals interacting with (a) technologies and (b) other individuals. Socio-technical interactions results in electronic traced data, which are referred to as social data. For instance, a Twitter user can tweet something, as where other Twitter users can join in on the topic by also tweeting, liking and retweeting. Therefore, these interactions generate various other micro-interactions (Vatrapu, 2010). Text data in the aspect of the theory relates to the structures and functions of social media platforms in terms of their technological intersubjectivity. Which describes the various ways in which people communicate with each other and influence each other through text (Vatrapu, 2016). Based on the theory of social data, Vatrapu (2016) presents a conceptual model of social data, which consists of interactions and conversation. The concepts of interactions includes the connection that forms from the usage of social media. The concept of conversations refers to the various parts of a social media interaction such as topics, emotions and keywords (Vatrapu, 2016). In this thesis, the theory of socio-technical interactions will be utilized to describe the dynamics of users' social media interactions on cryptocurrencies. In addition, Vatrapu's conceptual model of social data will be taken into consideration by investigating interactions and conversations among the users.

3.2.6 User-generated content in cryptocurrency

User-generated content can be broadly categorized into various forms, such as blogs, instant messaging apps, and social media platforms (Piñeiro-Chousa et al., 2022). These platforms can then create various types of content, such as photos, videos, and text. Many researchers have analyzed the role of user-generated content in the cryptocurrency domain. Mai et al. (2018) analyzed social media activity of Bitcoin users to predict the currency's future movements. The researchers also explored the economic impact of digital currency and social media. Grover et al. (2019) researched insights from user-generated content on Twitter relative to users' acceptance of blockchain technology. Chanson et al. (2020) studied the role of user-generated content in relation to blockchain and decentralized finance. Domingo et al. (2020) found that sentiment extracted from social media had a positive influences on ICO returns. Recently, Piñeiro-Chousa et al. (2022) studied the performance of DeFi coins in relation to other stocks, volatility, and user-generated content on Twitter. As mentioned, cryptocurrency and blockchain have gained much attention from researchers and academia in recent years. However, to the best of my knowledge, no study has directly analyzed the interaction of user-generated cryptocurrency content between larger coins contrary smaller coins.

3.3 Previous research

Previous research has shown that text data from Twitter can be used to view movements in the financial markets, especially within cryptocurrency. Despite the popularity of social media platforms such as Twitter, there is still a considerable amount of research being conducted on how investors can use them to predict the stock market's future performance. Online communities among users are also becoming more prevalent in the cryptocurrency space (Phillips and Gorse, 2017). Online communities can generate price predictions and trading signals by analyzing sentiment. Therefore, Twitter's relationship with the cryptocurrency markets has been studied in various papers (Phillips and Gorse, 2017).

Most studies have had the approach of using text data from Twitter to predict movements in the financial markets, especially within cryptocurrency. Lamon (2017) used social media and news sentiment to predict the price of cryptocurrencies. The model was trained using a classifier that learns feature weights for labeling data. A statistical model was able to predict the price of digital tokens in real-time. It was able to do so by analyzing both Twitter and news data. Tandon et al. (2021) aim to draw a correlation between the hyped tweets and the prices of cryptocurrencies. The paper aims to predict the future price of Bitcoin using the past values of the cryptocurrency. By analyzing various aspects of the project, such as financial data, the researchers drew a fine line between tweets' amount of impact on the market and the people who follow them. Kraaijeveld and De Smedt (2020) were able to find that the public sentiment data collected on Twitter could predict the returns of various cryptocurrencies such as Bitcoin and Litecoin. The researchers found Twitter sentiments to have predictive power through their statistical analysis.

Text analysis in social media is often used to study user behavior and trends. Researchers have observed that investors' sentiment on social media can influence reactions to news among users within a short period. Narman and Uulu (2020) studied the impact of positive and negative sentiments of social media users on cryptocurrency. The research aimed to understand the public opinion about cryptocurrencies by first understanding the dynamics of the users. This was conducted by applying a sentiment analysis. Alqaryouti et al. (2021) conducted a study on the impact of cryptocurrency usage on users' perceived benefits and behavior. The results of the study suggested that the perceived ease of use of cryptocurrency is associated with the usage behavior of digital currency.

According to behavioral economists such as Daniel Kahneman and Amos Tversky, decisions are based on both emotions and financial decisions. R. J. Dolan's work on sentiment analysis supports this idea, as it shows that demand for an asset can be influenced by more than just the economic fundamentals. The insights from these researchers open up the possibilities to find advantages through tools like sentiment and emotion analysis as it indicates that demand for an asset may be impacted by more than its economic fundamentals (Abraham et al., 2018). Karalevicius et al. (2018) conducted a study on the comments made by users on cryptocurrencies which revealed that the price of Bitcoin can influence the sentiment of users on social media platforms. The researchers also found that users tend to react quickly to news reports about the cryptocurrency.

3.4 Purpose and objective for the research

A person's opinions are often central to their actions and decisions. They are appraised by others based on how they see the world. The beliefs we have and our perception of reality is also influenced by how other people view the world (Liu, 2012). The rapid growth of sentiment and emotion analysis coincides with the rise of social media (Liu, 2012). For the first time in history, we have a huge volume of digital data that can be used to analyze people on platforms such as social networks, media, and forum discussions (Liu, 2012). The overall purpose of the research is to leverage social media to view interactions, sentiments, and emotions of tweets and users related to cryptocurrency.

This thesis is differentiated from the works outlined in the previous section in several key ways. Firstly, the aim is to perform a descriptive and exploratory case study focusing on the dynamic interactions among Twitter users. Secondly, data will be labeled on sentiments and emotions rather than actual price changes. Thirdly, the primary focus in this thesis will be on Bitcoin, Ethereum, Dogecoin and Uniswap, which are a combination of cryptocurrencies with larger market cap (Bitcoin and Ethereum), and cryptocurrencies with a smaller market cap (Dogecoin and Uniswap). In the past, previous academic research is mainly centered around Bitcoin and cryptocurrencies with larger market cap.

3.5 Knowledge gap and contribution

In the past, previous academic research has primarily focused on Bitcoin, its mentions, occurrence, and high price volatility. Little attention has been paid to cryptocurrencies such as DeFi coins and Meme coins in the context of social media analysis.

Because of the fast growth of digital currencies, the impact on social media platforms has become more prevalent. This is also evidenced by the rise of related news articles and videos. The impact users have on social media has continued to grow to the point where it has an impact on the consumer’s decisions on buying and selling. Cryptocurrency is growing at a fast pace and is distinct by not being backed by any physical commodities or assets. Sentiment and emotion analysis gives insight into user behavior, opinions and trends. Therefore, I consider it both interesting and important to explore.

The thesis will contribute to research on cryptocurrency in the context of social media analytics by studying the interactions generated from social media hype in larger cryptocurrencies contrary to smaller cryptocurrencies. The knowledge gap lies in exploring and comparing four different cryptocurrencies of both larger and smaller market cap, rather than only focusing on the larger coins such as Bitcoin and Ethereum. Sentiment and emotion analysis can help investors protect their financial interests by studying the public's opinions about cryptocurrencies. On the other hand, influence and effect of social media users are viewed. The thesis aims to study the scope of cryptocurrency in a data-driven approach by leveraging social media.

As described, the use of cryptocurrency is increasing, and its usability has gained attention from various perspectives. However, the research on cryptocurrency is insufficient, and there are rarely other currencies than Bitcoin that have been the subject of research. On that account, this research makes two main contributions. Firstly, as well as researching Bitcoin, the research will equally focus on Ethereum, Dogecoin, and Uniswap. Hence, the two largest coins by market cap, Bitcoin and Ethereum, and Dogecoin and Uniswap, which are smaller by market cap. Secondly, the focus is on the user perspective and their interactions with the different cryptocurrencies. The comprehensive aim of the research is to understand the dynamics of social media users’ interactions on cryptocurrencies by applying Social set analysis as framework. The Social set analysis is supported by sentiment analysis and emotion analysis for further analysis on users and events. Therefore the theoretical contribution is based on studying this gap by exploring interactions on cryptocurrencies among Twitter users, as well as contribution to literature on Social set analysis.

4 Methodology

This section will present the research methodology and analytical approach for the thesis. The methodology is divided into six parts. Each part will aim to provide valid and reliable methods to address the research aim and objectives. The methodology is benchmarked on relevant literature and research that will be further presented in the respective parts. The methodology consists of the following:

4.1 Research design: The research design will be presented as a multiple-case study approach.

4.2 Social set analysis: Social set analysis will be employed to understand the distributions and interactions between users across sets.

4.3 Sentiment analysis: The sentiment analysis identifies positive, negative and neutral sentiments.

4.4 Emotion analysis: The emotion analysis identifies the user's emotions.

4.5 Bot detection: Bot detection aims to detect the occurrence of Twitter bots in each dataset.

4.6 Event study: The event study aims to analyzes how user's interactions change before, during, and after a cryptocurrency-related event.

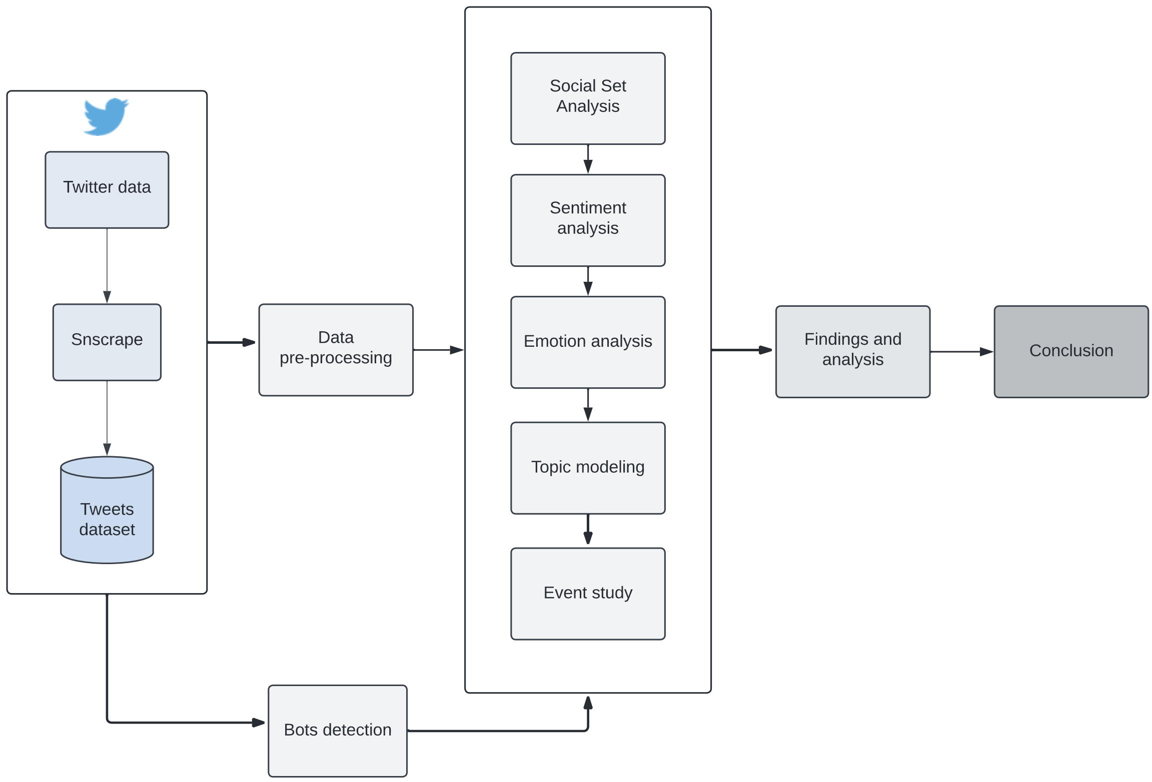


Figure 3. Overview of various phases in the thesis

4.1 Research design

Research design serves as the framework for developing effective research methods and techniques. It allows researchers to hone their skills in conducting studies relevant to the subject they study. It is a process utilized for carrying out various tasks and procedures in a research study. This involves guiding the researchers through the various steps involved in analyzing, reporting, and collecting data (De Vaus, 2006). It is also a strategy used to formulate and implement a study that effectively addresses a given research problem (De Vaus, 2006).

4.1.1 Multiple-case study approach

The strategy of the research in this paper is a case study. A case study is an in-depth examination of a specific case or cases where the context is events in the real world (Yin, 2009). A case study can narrow down a broad field of research or several researchable examples. A case study investigates the significance of real-life events (Yin, 2009). It can also shed light on a more significant case. Case studies are focused on an individual, group, or community (Yin, 2009).

A case study is characterized by typically focusing on depth rather than breadth; this means focusing on one instance of the phenomenon. The research consists of various sources and methods to collect data. In this thesis, a multiple case study will be conducted where each of the four cryptocurrencies represents a case.

Case studies are divided into three: Explorative, descriptive, and explanatory (Oates 2006, p. 162). An exploratory study is used to define a research question or a hypothesis used in a study. It can be used to identify topics that investigate real-life instances (Oates 2006, p. 143). The thesis explores different topics and issues related to a particular context. Therefore an exploratory study will be conducted. An exploratory study explores a particular topic in a particular context, in this case, Twitter users' interactions on cryptocurrency. As well as a descriptive study, that will analyzes social media data to identify user sentiments, emotions and topics.

4.2 Social Set Analysis

As earlier mentioned, the Social set analysis is a framework for analyzing the interactions and builds on set theory. The framework also combines social data's theoretical and practical aspects (Vatrapu et al., 2016).

Flesch et al. (2015) combined tools and methods for Social set analysis as a holistic approach for a case study. The researchers employed Social set analysis regarding the mobility of social actors across space and time. To uncover dynamic interaction among users over time and space, the researchers conducted set inclusions and exclusion of actors. Through the Social set analysis, the researchers were able to identify the various actor sets that are associated with marketing segmentations. These include brand loyalists, brand advocates, and social activists. Also Mukkamala et al. (2015) used the Social set analysis framework to study the structural properties of various corporate social media crises.

Vatrapu et al. (2016) utilized Social set analysis as a theoretical approach to big data analytics on case studies. The researchers collect social data from social media platforms to reveal interactional patterns. The researchers presented Social set analysis as a conceptual model of social data and a set-theoretical framework to analyze the massive amount of social data that is generated by social media platforms. The set formalization of the model provides a deeper understanding of the complex interactions generated by the social data. The study also applied a visual analytics dashboard to visualize the interaction of users. Vatrapu et al. (2016) refers to UpSet as a tool for combining visual analytics and set-theoretical data structures. In this thesis, Upset will be utilized to visualize sets.

Understanding the relationships between various sets of data is a fundamental part of data analysis. A set is a collection of elements that describe a common characteristic or shared opinions over the element (Lex et al., 2014). The concept of set visualization is used to explain the relationships between various data sets by identifying the co-occurrence of certain characteristics in a dataset (Lex et al., 2014). Lex et al. (2019) presented UpSet as a unique tool that combines the concept of set visualization and attributes visualization. While there are many sophisticated techniques for set-related tasks, there are also not enough feature-rich tools that can handle the complexity of these tasks (Lex et al., 2014). UpSet is compatible with a large number of sets, between 20-30 sets.

Gomez et al. 2020 used UpSet for survey data collected via complex sampling. The UpSet plots helped visualize the most commonly reported sets. Piccolboni et al. 2021 stated that the results of the research were visualized through an UpSet plot to show a graph's plot on specific rules, rather than the traditional Venn diagrams. In Beskow and Carley 2020, researchers used UpSet to study the differences between bots and human networks with data collected from Twitter. The UpSet graph was used supplementary to Venn diagrams to further explore the intersection of four sets.

In this thesis, Social set analysis is used to understand the dynamics and interactions between the different cryptocurrency coins. SSA is conducted by detecting the set of unique and common users and tweets in the different datasets. I will use UpSetPlot which is a Python library of UpSet plots by Lex et al. 2014. The plots are used to visualize set overlaps. The input format uses Python pandas. The series contains counts corresponding to each subset of the series. The index of the series shows the rows that are related to the categories (Nothman, 2021). Alteryx was used to format the data in correspondence with the requirements of the documentation provided by UpSetPlot. A snapshot of the Alteryx workflow is provided in appendix C.

4.3 Sentiment analysis

Sentiment analysis is a field of study that studies the various expressions of people regarding a certain topic, product, or organization (Liu, 2012, p.7). It is a commonly used analysis in Natural Language Processing (NLP). The expressions can be classified as positive, negative, or neutral. It uses various analytical techniques to analyze and compare the sentiments of different individuals. Indicators of sentiments are sentiment words and opinions words, which are used to express positive or negative sentiments. Examples of positive sentiments are words such as "good," "great," "wonderful", and "amazing," while examples of negative sentiments are words such as "bad", "horrible", "terrible" etc. (Liu, 2012, p.7).

Sentiments of users' posts can evaluate a person's social capital or signaling. By turning a private act of consumption into a public event, their sentiments can be used to identify their characteristics (Vatrapu et al., 2016). Through sentiment analysis, researchers can gain a deeper understanding of how people feel about the data being studied and further make informed decisions regarding the data (Sponder and Khan, 2018, p.277).

The use of sentiment analysis has been studied by various researchers in the field of information systems. In a study conducted by Pak and Paroubek (2010), the suggestion of separating tweets into neutral, positive, or negative categories would result in an efficient analysis of sentiments. A study conducted by O'Connor et al. (2010) revealed that the sentiment expressed in tweets related to several national polls of the public opinion. The researchers noted that sentiment analysis could be a cost-effective alternative to national polling, as it can accurately reflect the public's sentiments. Geetha and Bharathi (2017) studied the sentiment of the public on social media to predict the cryptocurrency market prices. Further, the researchers used an algorithm to analyze the data and establish the correlation between the sentiment and the market.

Sentiment analysis can be conducted by unsupervised lexicon-based models, supervised learning, machine learning models, deep learning etc. (Sarkar, 2019, p. 574). In this thesis, sentiment analysis will be implemented by applying lexicon-based and rule-based models. The purpose is to use the techniques to extract valuable data points from the unstructured text. The sentiment analysis aims to detect the positive, negative and neutral sentiments of tweets and users for further analysis.

The most common algorithms:

- **Rule-based models**

  Rule-based sentiment analysis is one of the very basic approaches to calculating text sentiments. It only requires minimal pre-work and the idea is quite simple, this method does not use machine learning to figure out the text sentiment (Shahul, 2021). For example: TextBlob and VADER use the bag-of-words approach where the text is considered to be the sum of its constituent words.

- **Word-embedding-based models:**

  In terms of text representation, embeddings are used in Natural Language Processing to represent similar words. These are represented using n-dimensional space vectors that are close to each other (Shahul, 2021).

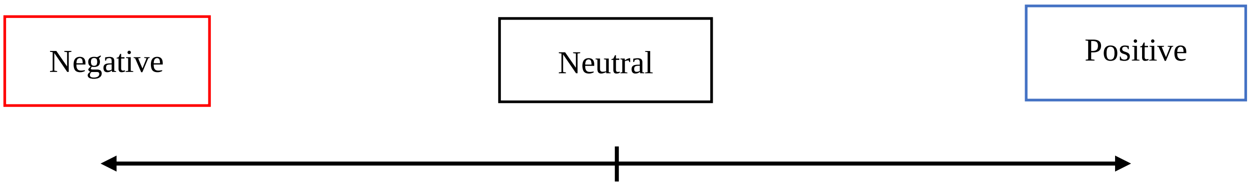


Figure 4. Sentiment analysis

4.4 Emotion analysis

The increasing number of conversations happening on social media has led to the ability to detect emotion and made possible a valuable strategic tool for the cryptocurrency industry. Over the past decades, various scientific streams and research groups have been developing theories and techniques related to emotions (Vornewald al., 2015). These studies have led to the establishment of a more general understanding of how humans behave. In terms of research, emotions have been regarded as a significant area of study within IS research. Various types of emotional constructs have been studied in studies related to the subjects (Vornewald al., 2015). Projections of feeling are the most visible layer of emotion in social media. They are often expressed through various forms of communication such as expressions and verbal statements on different social media platforms, for example on Twitter (Zimmerman et al., 2016). Understanding emotions is a complex task due to the varying characteristics of the user and the increasing number of internet slang terms (Chatterjee et al., 2019). Also, understanding the context and sarcasm in the text can lead to difficulties. Various studies on emotions have been conducted over the years.

Some of the most prominent studies on this topic include those conducted by Hochschild, 2002, Lane et al., 1996, Plutchik, 1994 (Chatterjee et al., 2019). Notable work in understanding and targeting emotions was also included in Ekman et al., 1983, where six class categorizations of emotions were conducted. In recent studies, Jussila et al. (2022) emphasize the importance of understanding the various emotions expressed in tweets to further understand the context of the interactions on social media. The researchers in Zimmerman et al. (2016) developed an intelligent method EmotionVis to detect the emotional expressions in Facebook posts. The tool works as a dashboard that gathers emotions from the text that is presented through visual analytics (Zimmerman et al., 2016).

There is a difference between sentiment analysis and emotion analysis. Unlike sentiment analysis, which merely divides data points based on a negative or a positive feeling, emotional analytics focuses on analyzing the psychology of consumers' (Kim and Klinger, 2018). However, the main target of an emotion analysis is to recognize the emotions in the text, rather than sentiment. This may make the task difficult because the difference in emotions is more distinct than the classification of positive, negative, and neutral (Kim and Klinger, 2018).

Emotions can be viewed as a finer-granularity sentiment. For example, anger and sadness can both be considered negative sentiments but are yet two different emotions. They can also be categorized as distinct feelings. This allows for an additional layer of analysis to understand the user's emotions. Therefore, in this thesis, it is interesting to not only look at the sentiment of the Twitter data, but to also view the emotions contained in the datasets.

There are various lists or models that can be used to identify the concept of basic emotions. In this analysis, emotions are based on the Python Library: Text2emotion. The Python package text2emotion helps extract emotions from the input content text2emotion (pypi.org, 2020). The library is compatible with five different emotion categories:

- Happy
- Angry
- Sad
- Surprise
- Fear

Aslam et al. (2022) used Text2emotion to study sentiment and emotion detection on cryptocurrency-related tweets. While communicating, humans use certain words and expressions in the appropriate context. The paper used Text2Emotion to identify the embedded emotion in the text and to produce a dictionary that can be used to interpret the data (Aslam et al., 2022)

4.5 Bots

A Twitter bot is an automated account controlled by a software program. It can perform tasks similar to those of a real Twitter user, but it's focused on specific goals(Johnson, 2020). The purpose of a bot can be beneficial or harmful, depending on its nature. For instance, it can be useful to broadcast weather emergencies in real-time. It can also be used to distribute informative content. A Twitter bot can also be designed to perform various malicious activities, such as spreading fake news campaigns and violating the privacy of other users (Johansen, 2020).

At Black Hat USA 2018, Duo Security researchers presented some of their work on tracking down Twitter bots (Johnson, 2020). The researchers built models that can predict the likelihood of bot-like activity on Twitter. They also used the Twitter API to build a data set that can analyze the behavior of different types of accounts (Lomas, 2018).

Reutzel (2018) highlighted common Twitter scams commonly used by bots and trolls. This includes tweets that state giving away free cryptocurrency in return for a small amount of money transferred, linking to other bots accounts related to cryptocurrencies which users will be asked to follow (Reutzel, 2018, cited in Kraaijeveld and De Smedt, 2020). The username of the account used by the bots is often used to impersonate other accounts. Or they also call other users and post a scam.

Despite the existence of Twitter bots, it is not considered a bad thing since our goal is to capture the market sentiment. In fact, as noted by Cresci et al. (2019), bots can have a semantic orientation and affect public sentiment. Machine learning techniques such as deep learning can remove bots from the dataset. However, a classifier machine learning bot classifier is beyond the scope of the research in this thesis. Therefore, existing literature will be used to detect typical characteristics found in twitter text from bots.

4.6 Event study

An event study is an empirical analysis that studies the impact of a major event on the value of a security, for example, a stock company (Hayes, 2022b). Time is the dependent variable in event analysis, it then looks for factors that explain the duration of the event or the time until the event occurs (Hayes, 2022b). The methodology of an event study is widely used to view the effects of events in the financial markets (Kliger & Gurevich, 2014). The concept of the event study emerged from studies conducted by Ball and Brown (1968) and Fama et. al (1969) (Bowman, 1983). (Ball and Brown (1968) conducted an event study to investigate security price reaction regarding annual accounting earnings. Fama et. al (1969) studied how security prices change as new information is released. Rather than following a market-focused approach, the researcher viewed how prices respond to changes in the market (Bowman, 1983). In recent years, also social media data has been leveraged to conduct event studies. Flesch et al. (2015) conducted an event study to investigate the reflections of real-world events regarding user interactions on social media platforms.

There isn't a unique methodology for an event study, however the aim is to identify three important time periods (Menon et al., 2018). It takes into account three important time periods: the event, the pre-event, and the post-event window (MacKinlay, 1997, cited in Menon et al. 2018). The first step in the event study process is to identify the event's interest and the event window which is the period of time when the event is active (Menon et al., 2018). Further, the pre-event window is defined as the period before the event. Thirdly, the post-event window is the time after the event window (Menon et al., 2018).

The event study methodology was applied to identify user interactions in social media before, during and after the event. Menon et al. (2018) used Facebook data to understand the interactions between investors and crowdfunders before, during, and after a crowdfunding campaign. A definition of events for each of the four cryptocurrency coins, as well as defining the event timeline will be presented in the "Analysis and findings" part of the thesis. The event study in this paper will be conducted to investigate sentiments, emotions, and topics before, during, and after an event has occurred to analyze change over time and in the face of different events.

5 Data

In this section, the data for the thesis will be presented. Firstly, the data collection will be presented where data extracting using Python library Snscrape is described. Followed by data pre-processing and emoji handling.

5.1 Data collection

The data collection process consisted of extracting data by scraping tweets off Twitter with Python library Snscrape to create datasets. The data collected for the thesis was for the cryptocurrencies Bitcoin, Ethereum, Uniswap and Dogecoin. The datasets were created separately by scraping tweets off Twitter with Python library Snscrape with keywords for each cryptocurrency. Snscrape was released on July 8, 2020, and is a tool that enables users to scrape data from social networks (JustAnotherArchivist, 2020). It scrapes content such as users, hashtags, threads, lists and posts without using Twitter's API. Snscrape requires Python 3.8 or higher.

There are various methods of scraping data from Twitter, among the most popular are also Tweepy and Twint. Tweepy is a Python library for accessing Twitter API. It is a great tool for creating simple automation, but it has a limit of 3200 tweets and a time limit of seven days (tweepy.org). Additionally, there is no access to historical data. Twint is a Python-based Twitter scraping tool that allows you to scrape tweets from various Twitter profiles without using Twitter's API. It does so by extracting Tweets related to specific topics, hashtags, trends and users' (Twint, 2021). Twint has no limit on the amount of tweets that can be scraped. However, I found Twint to be challenging because of the minimal documentation. Executing codes that weren't in the example code required orders of magnitude. Additionally, I had problems with consistency, where the set maximum parameter wasn't always returned.

On the other hand, the Snscrape library is an excellent example of a library that allows users to extract data from social media easily. It doesn't require an API and has no limit on how much data it can pull. Snscrape can return thousands of tweets within minutes using the text search query, hashtags or user profiles. Snscrape can be used in two ways: command prompt/terminal or Python wrapper (Desai, 2022). I used the Python wrapper to scrape data because I believe it is easier to interact with. In research, Snscrape was used by Ridhwan and Hargreaves (2021) to extract tweets to leverage Twitter data to understand public sentiments. Sarkar and Rajadhyaksha (2021) utilized Snscrape to extract tweets in their analysis of Twitter linguistics.

Tweets were collected separately for each of the cryptocurrencies between the time period January 1st 2020 till January 31th 2022. The defined parameter was set to 20000 tweets per month throughout the time period, additionally only tweets in English were scraped and fetched. Therefore, all non-English tweets were filtered out to avoid having a mixed-language data set. Each tweet contained DateTime, tweet ID, text and username. Due to the amount of Twitter data extracted, the scraping had to be conducted in batches. The scraped data was stored in CSV files and merged into one final CSV file per cryptocurrency. This resulted in four datasets, one for each of the cryptocurrencies. The keywords for Ethereum were “Ethereum” and “ETH”, returning in total 500000 tweets in the time period. The keywords for scraping Bitcoin tweets were “Bitcoin” and “BTC”, returning in total 500000 tweets in the time period. The keyword for Dogecoin was initially “Dogecoin” and “Doge”, but this only returned 1800 tweets in January 2020. The keyword was then changed to only “Dogecoin” and it then returned 4578 tweets in January 2020. Dogecoin was a fairly new coin in this time period. The keyword Doge in this period mostly returned automatic price updates from bots containing the #Doge. When only using the keyword “Doge”, it returned 20000 tweets as requested, but all the content was not crypto-related. In contrast, the keyword “Dogecoin” returned more tweets than the keywords “Dogecoin” and “Doge” in 2020. In total, 86049 was scraped with only the keyword “Dogecoin”, and 35060 with both keywords “Dogecoin” and “Doge”. When scraping the sample for Dogecoin tweets from 2021, both the keywords “Dogecoin” and “Doge'” were used. This returned 20000 tweets per month as requested. Towards the end of 2020, Elon Musk tweeted continuously about Dogecoin which caused the price to increase by 25% on 20th December. He simply tweeted one word: “Doge”. According to Google trends, Dogecoin increased in popularity in January 2021 and peaked in May 2021. Thus 86049 tweets were scraped for Dogecoin in 2020, and 260000 for 2021 (including tweets for January 2022), in a total 346049. The keywords for Uniswap was “Uniswap”, returning in total 378249 tweets. 118279 tweets were scraped in 2020. 260000 tweets were scraped in the time period January 2021 to January 2022. Fewer tweets were scraped in 2020 due to the Uniswap token not being launched before September 17th 2020 (CoinMarketCap, 2022). It was first in August 2020 and beyond that 20000 tweets per month containing Uniswap were scraped. The process resulted in a total of four datasets with a total of 1724328 public Tweets.

	Bitcoin	Ethereum	Dogecoin	Uniswap
2020	240 000	240 000	86 049	118 279
2021	260 000	260 000	260 000	260 000
Total dataset	500 000	500 000	346 049	378 279

Table 2. Overview over scraped data. 2021 includes January 2022

5.2 Data pre-processing

Data preprocessing involves techniques that are used to prepare the data for further analysis. Before the actual analysis begins, the preprocessing stages are carried out to prepare the raw data. Pre-processing helps transform the data into a clean and useful set and involves steps such as cleaning, transforming and adapting the data for the next steps in the analysis process. Data preprocessing is important because raw tweets are often cluttered and contain much noise. Twitter data is complex and lacks structure. Therefore, Twitter data requires extensive preprocessing to extract its full potential.

I used the Python library Preprocessor, a preprocessing library for tweet data written in Python (pypi.org). The library makes it easy to clean, parse or tokenize tweets. Also, natural language processing (NLP) techniques were used to clean the text data. Further, I followed a similar pre-processing approach used in Kraaijeveld and De Smedt (2020).

First, normalization was applied to the text data by removing URLs, http/https, extra white spaces and user mentions (@). Elements that don’t add value to the meaning of the tweet, such as punctuation marks and special characters, hyperlinks, hashtags, Twitter handles and metadata (‘RT’ for retweeted) were removed. Special characters and symbols are generally non-alphanumeric characters that add to the noise in the text (Sarkar, 2019, p 138). Nevertheless, emojis were not removed from the dataset. Emojis in the text were translated and converted into the Unicode text equivalents. Emoji handling will be discussed in the next part.

For the removal of stop words, the genism library was applied. Examples of removed stop words are: “the”, “is”, “and”, “to”, “or”, etc. The stop words don't have value in a sentiment analysis because stop words don't indicate positive or negative sentiments (Kraaijeveld and De Smedt, 2020). For this reason, they are removed from the dataset.

Lastly, all duplicate rows were removed to be left with only unique tweets. Duplicates were removed in case of the same tweet being posted multiple times or scraped multiple times. The dataset contained a large amount of duplicates.

I explored stemming the words by using the NLTK Python package PorterStemmer. Its effect was too aggressive; therefore, stemming of the words was not included in the final dataset (Kraaijeveld and De Smedt, 2020). All numerical numbers were also removed from the datasets. As a sanity check, I frequently printed the pandas DataFrame to a CSV file in excel and checked that the executed pre-processing and changes had been completed.

Data cleaning and preprocessing were adapted to the structure of the VADER and Textblob sentiment analysis lexicons that will be further presented in the analysis. For example, I did not transform letters to lower case because VADER takes capital letters into consideration to emphasize sentiments (Hutto and Gilbert, 2014).

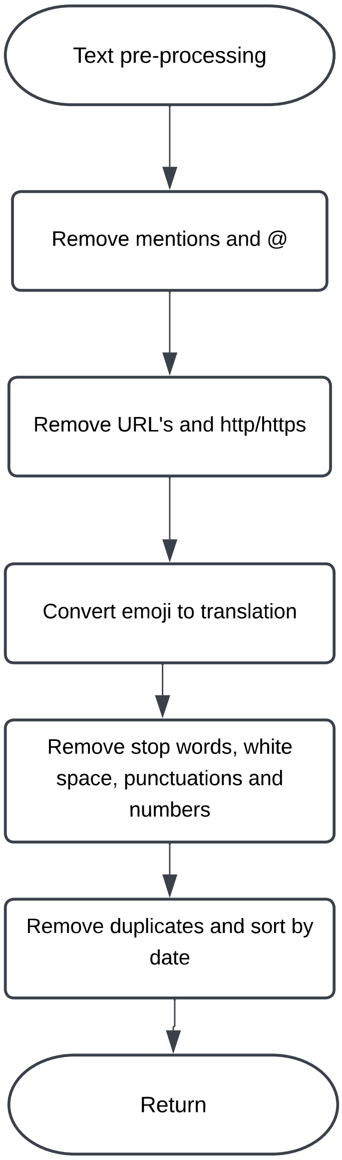


Figure 5. Steps in text pre-processing

Processing technique	**Result**
Original Tweet	@Cryptolover #Ethereum has risen 63% for the month of April😱, pushing it to the front. Definitely sitting at the top of the range but it has been steady price action so far. #cryptocurrency $ETHUSD $ETH #Crypto #cryptotrading https://twitter.com/
Translate emoji to word	@Cryptolover #Ethereum has risen 63% for the month of April shock, pushing it to the front. Definitely sitting at the top of the range but it has been steady price action so far. #cryptocurrency $ETHUSD $ETH #Crypto #cryptotrading https://twitter.com/
Remove URLs	@Cryptolover #Ethereum has risen 63% for the month of April shock, pushing it to the front. Definitely sitting at the top of the range but it has been steady price action so far. #cryptocurrency $ETHUSD $ETH #Crypto #cryptotrading
Remove whitespace white space, mentions and special characters	Ethereum has risen 63 for the month of April shock, pushing it to the front. Definitely sitting at the top of the range but it has been steady price action so far. cryptocurrency ETHUSD ETH Crypto cryptotrading
Remove numbers	Ethereum has risen for the month of April shock, pushing it to the front. Definitely sitting at the top of the range but it has been steady price action so far. cryptocurrency ETHUSD ETH Crypto cryptotrading
Remove stop words	Ethereum risen month April shock, pushing front. Definitely sitting top range steady price action far. cryptocurrency ETHUSD ETH Crypto cryptotrading

Table 3. Example of pre-processed tweet

Cryptocurrency	**Before pre-processing**	**After pre-processing**
Bitcoin	500 000	349 654
Ethereum	500 000	320 688
Uniswap	378 279	291 051
Dogecoin	346 049	230 683

Table 4. Number of Tweets before and after pre-processing

The final output of the pre-processing step resulted in four datasets with a total of 1192076 number of public Tweets.

5.2.1 Emoji handling

Along with the text, there are also other high-impact forms of expression, such as emojis. These are simple and powerful tools used to communicate. Often tweets and posts on social media are followed by an emoji. The function of an emoji is to fill in emotional cues that otherwise are missing from the typed text.

The dataset contained a large amount of emojis. The approach of handling emojis was to convert emojis into words in order to preserve the emojis' information. The emojis were converted according to Unicode.org, the world standard for text and emoji. The translation converts all emojis in the text to the Unicode text equivalents. In Boukhers et al. (2022), all the emojis in the dataset were kept and converted to the translation of Unicode. A study by Arifiyanti and Wahyuni (2020) on tweet sentiment classification on Indonesian Twitter users claims that due to the increased use of smartphones, 45% of the tweets in the dataset contain emojis, which affects the sentiments of the tweets. Therefore, all emojis were converted in translation with the Unicode translation. Because users use emojis when tweeting to convey their message efficiently and express opinions and sentiments about topics, converting the translation was considered in this paper.

Emoji	Unicode translation	Converted translation
😌	Relieved face	Relieved
😃	Grinning face with big eyes	Grinning
🤥	Lying face	Lying
🤮	Face vomiting	Vomiting
😦	Frowning face with open mouth	Frowning

Table 5. Examples of emoji conversion

Specific emojis used in correlation with cryptocurrency in social media by the cryptocurrency community are translated. Emojis are a big part of the language of investing and discussion of cryptocurrencies.

Emoji	Unicode translation	Converted crypto translation
🚀	Rocket	Increase
🦍	Gorilla	Strong
💎	Diamond	Hold
🐳	Whale	Volume

Table 6. Converted cryptocurrency specific terms

The rocket emoji means that the price of a cryptocurrency has reached its peak value and is rising off the charts, the term “to the moon” is also used in correlation with this emoji, meaning that there’s a quick rising in price. A whale emoji is used to describe a person who has enough coins or tokens to affect the prices of various commodities in the market significantly. Whales are known to cause substantial buy orders on the market, which can raise the price of a cryptocurrency (Mitra, 2021). The diamond emoji symbolizes investors who hold on to their assets, whether they’re winning or losing at the time being. The gorilla emoji implies the saying “apes together strong”. This saying was first used by the Redditt community WallStreetBets, who used the term to refer to investors who are bullish on heavily-shorted assets (Hartwig, 2021). Wang and Luo (2021) found various examples of self-referential humor and esoteric slang in the users' language in a study of stock price and sentiments in social media. The researchers found heavy use of emojis was influenced by the terms used by the WallStreetBets community. The researchers replaced emojis with descriptive text tags. Mahmoudi et al. (2020) performed a comprehensive study where domain-specific emojis in sentiment classification were identified. The study provides examples of cryptocurrency specific emoji’s translation. For example that the rocket emoji (🚀), transfers the meaning of an increase in price.

In this paper, the descriptive text tags have been conducted in correlation to cryptocurrency-specific terms used by communities such as Redditt WallStreetBets. Therefore, the converted cryptocurrency-specific terms have been managed in correlation with previous research and cryptocurrency glossary and specific terms.

An important factor when converting the emojis in accordance to the cryptocurrency communities' translation is to convert this in a way that the lexicon can read and interpret the meaning. Emojis that don’t give any implication of sentiments were removed. For example clothing, musical instruments, geometrics, country flags, zodiac signs, and gender.

6 Analysis and Findings

In this section, the analyses will be conducted and presented through the methodological approaches. The analysis is benchmarked on social set analysis. In the social set analysis, sentiment and emotion analysis will serve as a backdrop for further analysis on users and event study.

6.1 Sentiment Analysis of Tweets

6.1.1 VADER

VADER is a sentiment analysis tool that focuses on social media sentiments, developed by C.J. Hutto. VADER stands for Valence Aware Dictionary and sEntiment Reasoner (Sarkar, 2019, p.584). It uses a combination of a sentiment lexicon and a rule-based sentiment analysis system that can analyze and interpret the various words and phrases commonly used in social media. It can do so without requiring any training. In addition to giving scores on positivity, negativity and neutrality, VADER also indicates how positive or negative a sentiment is. Abraham et al. (2018), Kraaijeveld and De Smedt (2020), and Narman and Ulu (2020) all performed a VADER sentiment analysis with data extracted from Twitter data to study social media's impact on cryptocurrency.

VADER identifies generalizable heuristics humans use to assess sentiment intensity in the text (Hutto and Gilbert, 2014). Capitalization increases the intensity of a word's sentiment. For example, capitalizing the word GREAT makes it more pronounced. Conjunctions such as "because" and "but" signal a shift in the sentiment. For instance, “the food here is bad, but the service is excellent” has mixed sentiment, with the ending dictating the overall rating (Hutto and Gilbert, 2014). In addition, punctuations such as the exclamation point increase the sentence's intensity.

The VADER lexicon contains over 7,500 lexical features with validated valence scores that indicate the sentiment polarity (Sarkar, 2019, p.584). The lexicon is based on a dictionary that assigns predetermined sentiment scores measured on a scale from -4 to +4, where -4 is the most negative and +4 is the most positive, while 0 represents neutral (Hutto and Gilbert, 2014). Further, VADER generates a valence-based lexicon with focus on the intensity and polarity of the sentiments. The compound score of VADER is a numerical representation of the intensity of the input text. It considers the feature's valence score, which is adjusted according to the rules, and then normalized to be between -1 which is the most extreme negative, and +1 which is the most extreme positive. The negative, neutral, and positive scores are the ratios for the proportions of text within each category. They are the most useful metrics for analyzing the presence of sentiment in a given sentence (Hutto and Gilbert, 2014).

VADER	Bitcoin	Ethereum	Uniswap	Dogecoin
Post tweets	**349 654**	**320 688**	**291 051**	**230 683**
Positive	169 647	161 964	164 444	127 388
Negative	64 818	49 407	41 430	29 103
Neutral	115 189	109 317	85 177	74 192

Table 7. Results from VADER Sentiment Analysis

Tweet	**Compound**	**Sentiment**
“Dear Bitcoin its time increase bullish”	Neg: 0.0, Neu: 0.505, Pos: 0.495, Compound: 0.5994	Pos: 0.5994
“Why Bitcoin usability still so bad decade later”	Neg: 0.202, Neu: 0.798, Pos: 0.0, Compound: -0.5849	Neg: -0.5849
“Bitcoin Art Near Downtown Dallas”	Neg: 0.0, Neu: 1.0, Pos: 0.0, Compound: 0.0	Neu: 0.0

Table 8. Example of tweet sentiments. Neg stands for negative tweets, Pos for positive, and Neu for neutral tweets. Compound is the normalized results of the three polarity scores.

6.1.2 TextBlob

TextBlob is a Python library that supports Natural Language Processing (NLP). It was developed using the NLTK framework, which is a set of tools that allow users to perform various tasks related to the analysis and classification of text (TextBlob, 2020). TextBlob supports complex analysis and operations on text data. Aslam et al. (2022) used TextBlob to label cryptocurrency-related tweets with sentiments. In a study by Gurrib et al. (2021), TextBlob was used to determine the sentiments and polarity of tweets related to Bitcoin price.

A sentiment is defined by the semantic orientation and the intensity of the words and stenches when there is a lexicon-based approach. This requires a pre-defined dictionary that classifies which words that are negative and which that are positive. A sentiment is defined by its orientation and the intensity of the various words in a sentence. The final sentiment is computed by taking an average of all the scores.

The TextBlob sentiment analyzer returns polarity and subjectivity of a sentence. Polarity is a lies between [-1,1], where -1 indicates negative sentiment and +1 indicates positive sentiments (TextBlob, 2020). Subjectivity measures the amount of factual information and personal opinion in the text and the range is [0,1] (TextBlob, 2020). Words of negation have the effect of reversing the polarity. The library also has semantic labels that strengthen the analysis, such as emojis and exclamation marks (Sarkar, 2019, p. 557). Another parameter of TextBlob is intensity, which is used to determine if a word modifies the next word. For example, adverbs such «very well» are used as modifiers.

Textblob	Bitcoin	Ethereum	Uniswap	Dogecoin
Post tweets	**349 654**	**320 688**	**291 051**	**230 683**
Positive	146 730	149 731	140 074	89 847
Negative	48 689	44 317	34 412	23 191
Neutral	154 235	126 640	116 565	117 645

Table 9. Results from Textblob Sentiment Analysis

Tweet	Compound	Sentiment
“Bitcoin Art Near Downtown Dallas”	0.1	Positive
“Why Bitcoin usability still so bad decade later”	-0.349	Negative
“Dear Bitcoin its time increase bullish”	0.0	Neutral

Table 10. Example of tweet sentiment TextBlob (Bitcoin)

6.1.3 Comparing results

Previous research and literature such as McDonald and Loughran (2011) observed that the performance of a sentiment analysis improves its performance when using a specific dictionary or lexicon. The results presented the efficiency of contextual embeddings in sentiment analysis based on the comparison of lexicons and fixed word and sentence encoders.

VADER	Positive	Negative	Neutral	***TextBlob***	Positive	Negative	Neutral
Bitcoin	169 647	64 818	115 189	**Bitcoin**	146 730	48 689	154 235
Ethereum	161 964	49 407	109 317	**Ethereum**	149 731	44 317	126 640
Uniswap	164 444	41 430	85 177	**Uniswap**	140 074	34 412	116 565
Dogecoin	127 388	29 103	74 192	**Dogecoin**	89 847	23 191	117 645

Table 11. Results from Sentiment Analysis

The results show that the VADER lexicon returns more positive sentiments than TextBlob, while TextBlob returns more neutral sentiments than VADER. TextBlob also returns fewer negative sentiments than VADER.

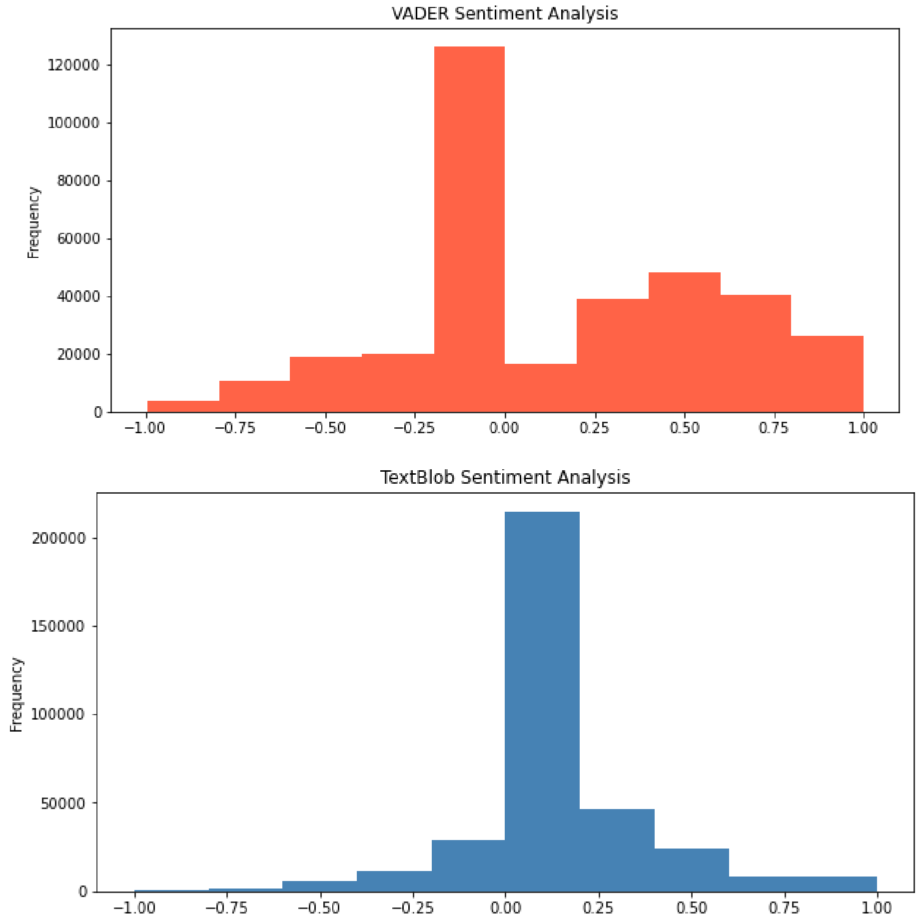


Figure 6. Example of Bitcoin VADER and TextBlob sentiment scores

6.1.4 Evaluation metrics

Classification models are usually measured on how well they predict the outcome of new data points. One of the easiest ways to measure the accuracy of a sentiment analysis is by performing an accuracy test. This allows the model to see how many times it correctly resulted in a given outcome. As well as precision, recall, and F1-score (Sarkar, 2019, p.309). Hutto and Gilbert 2014 also stated that they rely on three metrics to quantify the VADER tool's ability to classify sentiments: precision, recall, and the F1 score. The evaluation metrics have been carried out according to Sarkar's module for model evaluation, model_evaluation_utils in Python. The module leverages the Scikit-Learn metrics module to compute most of the evaluation metrics and plots (Sarkar, 2019, p.309)

Accuracy defines how correctly the model predicts something. It is also known as the overall proportion of the correct predictions in the model (Sarkar, 2019, p.313). The formula for executing accuracy is the following:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Precision is a positive predictive of how many predictions are correct (Sarkar, 2019, p.313). The formula for precision is the following:

$$Precision = \frac{TP}{TP + FP}$$

Recall is used to measure the percentage of data points that were correctly predicted (Sarkar, 2019, p.313). It is also known as the hit rate, coverage, or sensitivity metric. The formula for conducting recall is the following:

$$Recall = \frac{TP}{TP + FN}$$

The F1-Score aims to harmonize the mean of precision and recall in the cases where a balanced optimization is wanted (Sarkar, 2019, p.314). The formula is the following:

$$F1\,Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

VADER

Metrics	Bitcoin	Ethereum	Dogecoin	Uniswap
Accuracy	51.2	51.2	57.9	56.2
Precision	53.8	54.2	60.9	59.7
Recall	51.2	51.2	57.9	56.2
F1 Score	47.2	48.0	55.8	53.7

Table 12. Model Performance Metrics for VADER

Textblob

Metrics	Bitcoin	Ethereum	Dogecoin	Uniswap
Accuracy	50.0	55.1	45.6	55.0
Precision	45.0	50.1	40.1	50.4
Recall	50.0	55.1	45.6	55.0
F1 Score	43.2	49.1	39.7	49.5

Table 13. Model Performance Metrics for Textblob

From the model performance metrics, VADER overall has higher metric scores. In the performance review by Stenqvist and Lönnö (2017), VADER and eleven other sentiment analyses were compared. The result showed that VADER outperformed other techniques in the social media text domain. Therefore, VADER will be used as the baseline for sentiment classification in further analysis.

6.2 Emotion analysis

In this part, I will analyze and discuss the emotions in the datasets. After collecting 1724328 million tweets, and utilizing 1192076 million tweets after preprocessing and removing duplicates. Each dataset with tweets of Bitcoin, Ethereum, Dogecoin, and Uniswap will be classified individually. The five different emotion categories in the Text2emotion library are the following:

- Happy
- Angry
- Sad
- Surprise
- Fear

Bitcoin

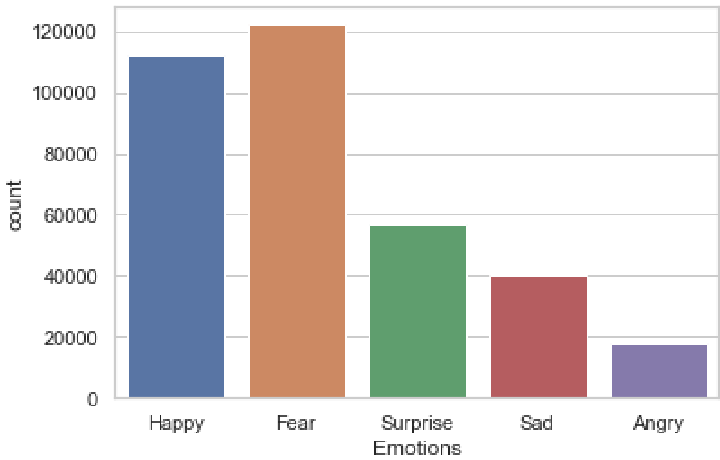


Figure 7. Emotion analysis distribution Bitcoin

The top emotion in the Bitcoin dataset is fear, with a frequency of 122032 (35%). Followed by the emotions happy (32%), surprise (16%), sad (12%), and angry (5%).

Ethereum

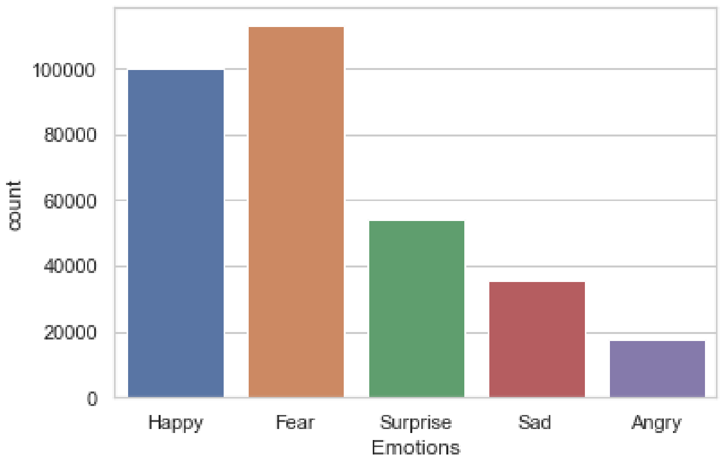


Figure 8. Emotion analysis distribution Ethereum

The top emotion in the Ethereum dataset is fear, with a frequency of 113088 (35%). Followed by the emotions happy (31%), surprise (17%), sad (11%), and angry (6%).

Uniswap

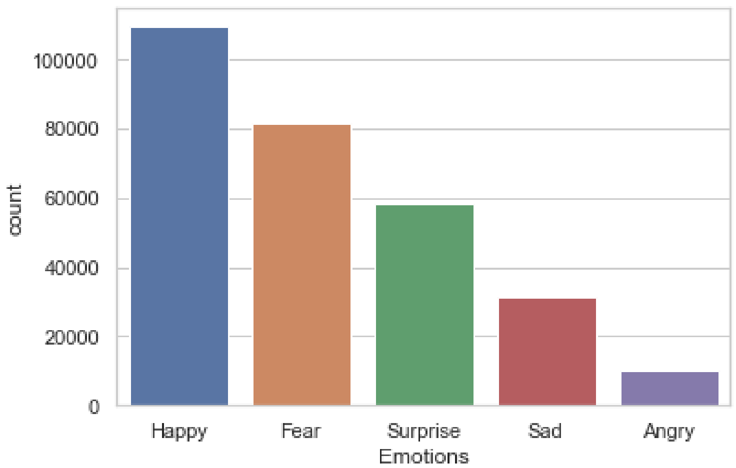


Figure 9. Emotion analysis distribution Uniswap

The top emotion in the Uniswap dataset is happy, with a frequency of 109555 (38%). Followed by fear (28%), surprise (20%), sad (10), and angry (4%).

Dogecoin

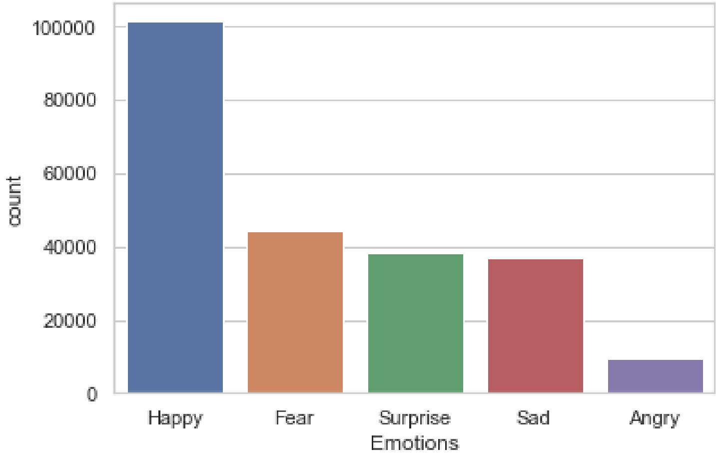


Figure 10. Emotion analysis distribution Dogecoin

The top emotion in the Dogecoin dataset is happy, with a frequency of 101 529 (44%). Followed by fear (19%), surprise (17%), sad (16%) and angry (4%).

	Happy	Fear	Surprise	Sad	Angry
Bitcoin	112 183	122 032	56 899	40 422	18 118
Ethereum	99 924	113 088	54 173	35 637	17 866
Dogecoin	101 529	44 316	38 227	37 089	9 522
Uniswap	109 555	81 617	58 171	31415	10 293

Table 14. Identified emotions

6.3 Descriptive analysis

In this part, the descriptive analysis will visualize data points, present mean values, how scattered the datasets are and how the points in the data are shaped. Further, a word frequency analysis will be provided. The various graphs and visualizations will utilize color theory where the colors are in the respective colors of each cryptocurrency's logo. Color theory refers to the science and art of using color.

6.3.1 Descriptive statistics

All data were collected between January 1st 2020 and January 31st 2022. The largest volumes of data are concentrated between August 2021 and November 2021.

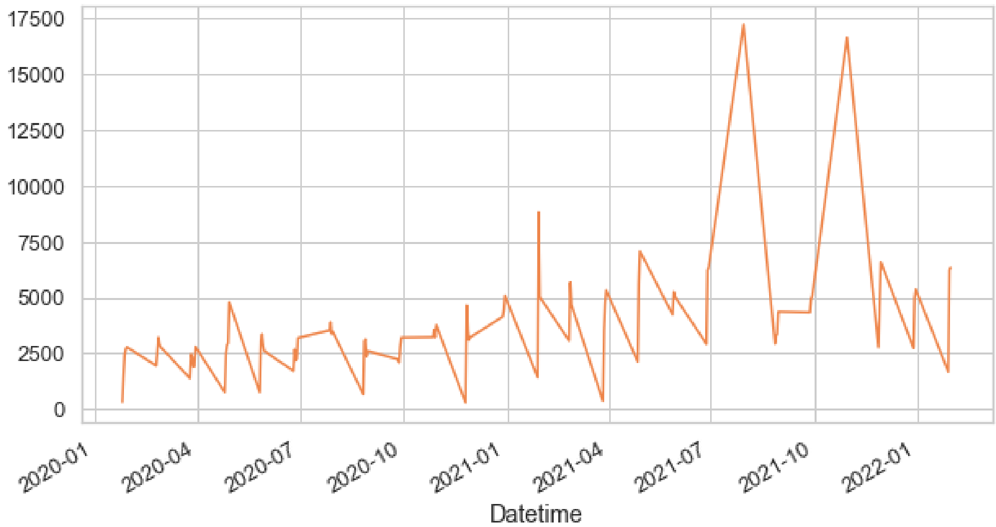


Figure 11. Bitcoin distribution graph

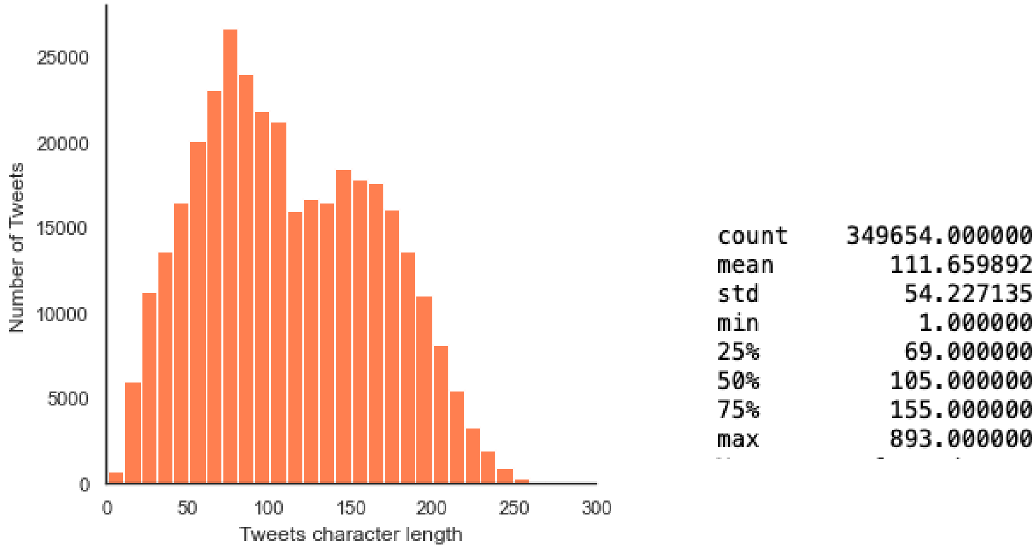


Figure 12. Bitcoin distribution of Tweet character length

The Bitcoin distribution is right skewed which indicates that majority of the Bitcoin tweets are short in text.

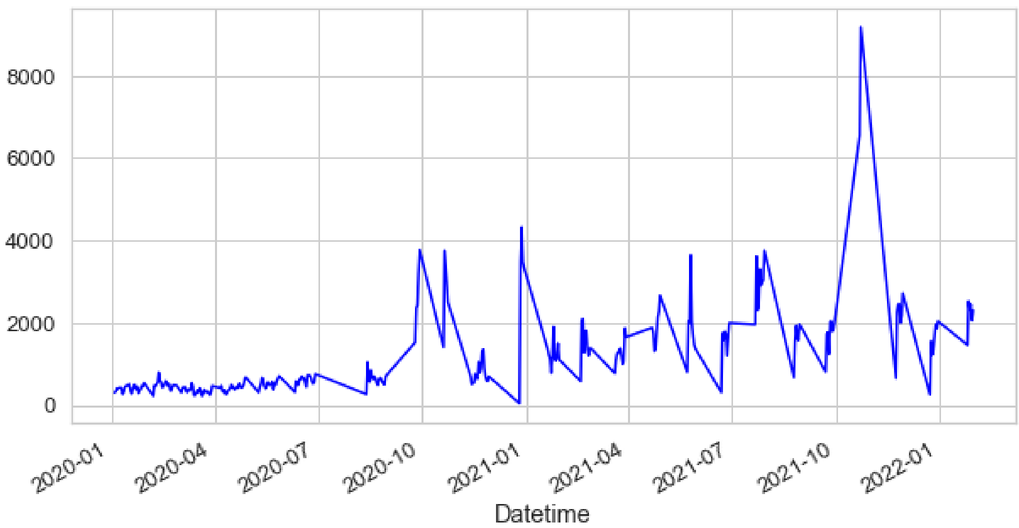
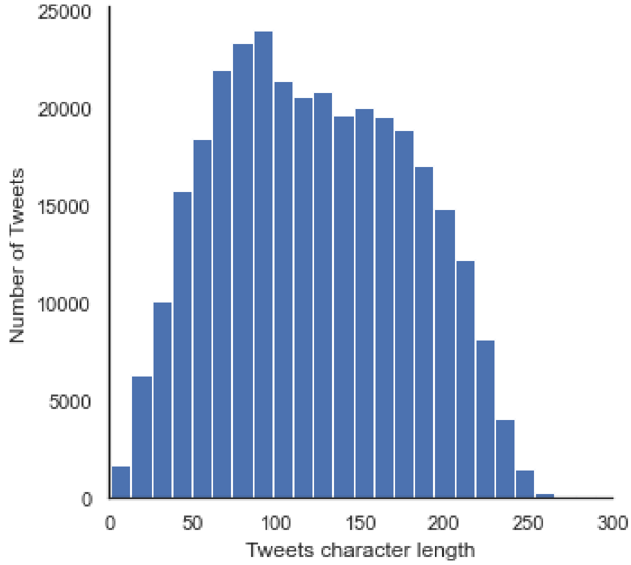


Figure 13. Ethereum distribution graph

The largest volumes of data are concentrated at November and December 2021.




```
count    320688.000000
mean         17.679754
std           7.953488
min           1.000000
25%          11.000000
50%          17.000000
75%          24.000000
max         123.000000
```

Figure 14. Ethereum distribution of Tweet character length

The Ethereum distribution has a normal distribution, which indicates that the tweets have combination of longer and shorter text.

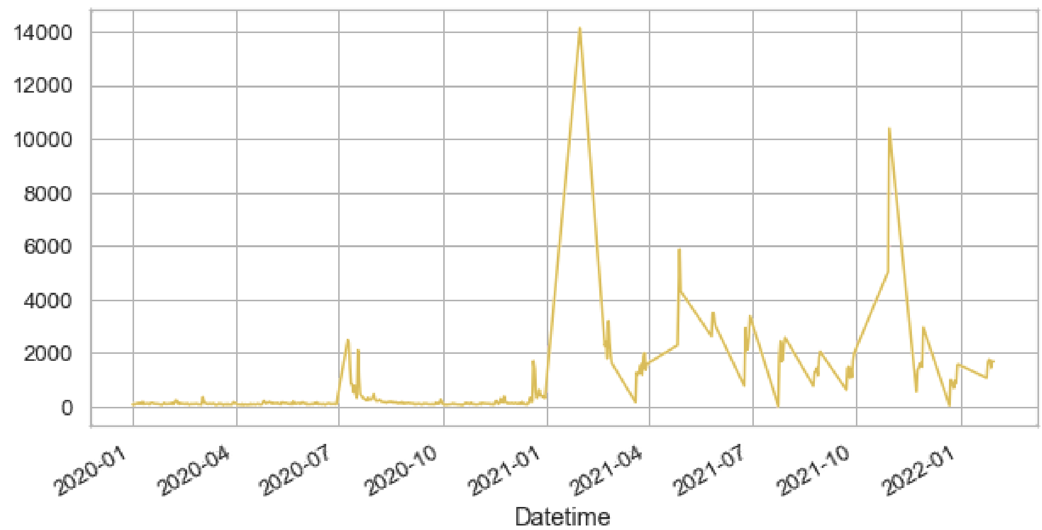
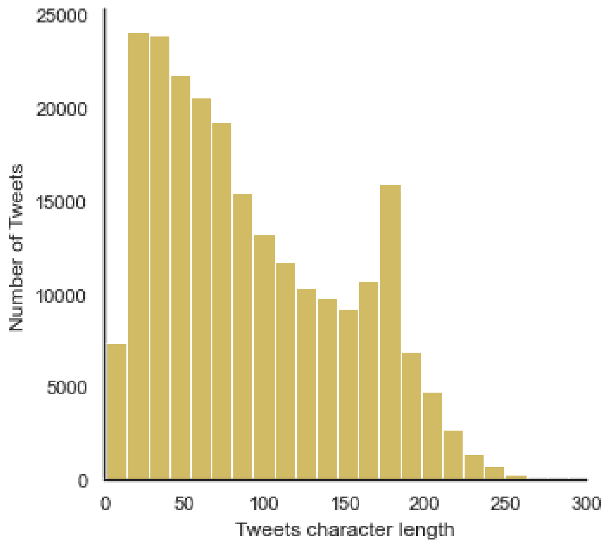


Figure 15. Dogecoin distribution graph

The largest volumes of data are concentrated between January-April 2021. As well as November 2021 to December 2021.




```
count    230683.000000
mean         92.028515
std          59.466979
min           1.000000
25%          42.000000
50%          78.000000
75%         138.000000
max        1178.000000
```

Figure 16. Dogecoin distribution of Tweet character length

The Dogecoin distribution is right skewed which indicates that majority of the Dogecoin tweets are short in text.

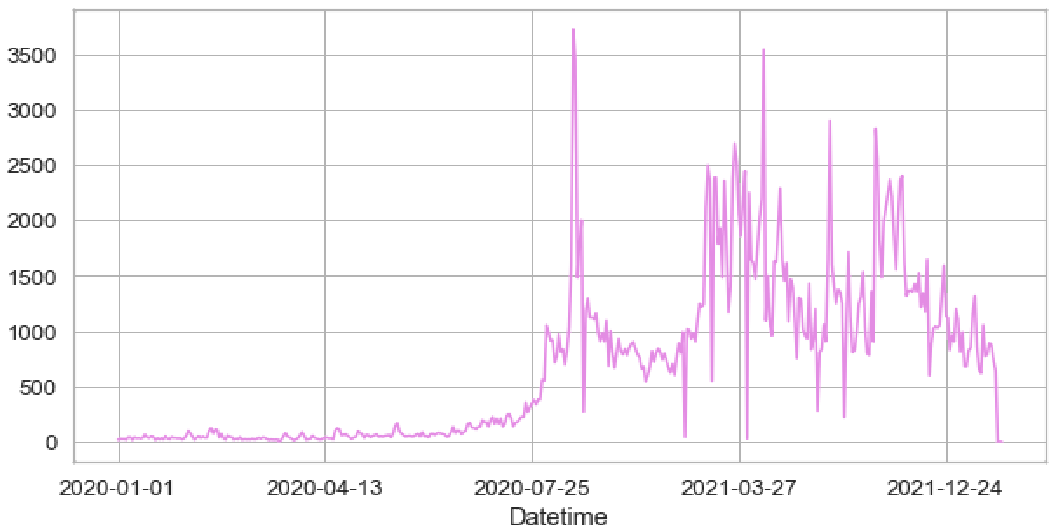


Figure 17. Uniswap distribution graph

The largest volumes of data are concentrated at July 2020, as well as April 2021.

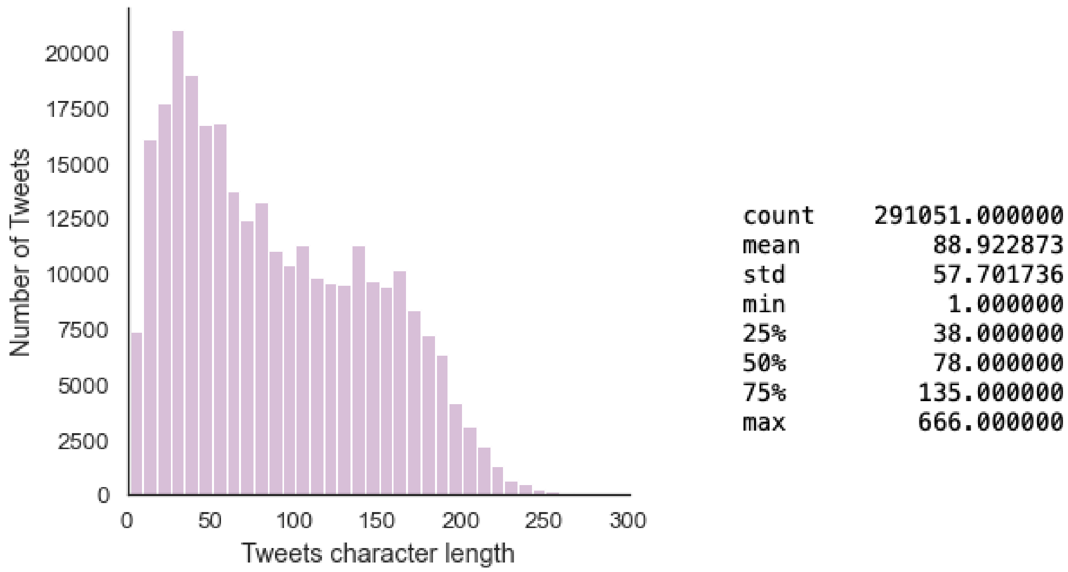


Figure 18. Uniswap distribution of Tweet character length

The Uniswap distribution is right skewed which indicates that majority of the Uniswap tweets are short in text.

Cryptocurrency	Max word count	Max word length	Mean for wordcount	Average word length
Bitcoin	149	893	16	111
Ethereum	123	1084	17	122
Uniswap	83	666	14	89
Dogecoin	133	1178	13	94

Table 15. Word count and length of in the cryptocurrency datasets

The tweets have an average word count of between 13-16 words, which is quite similar. The average length however varies slightly, but not significantly. Thus, the length of the tweets is not a strong indicator of polarity.

6.3.2 Word frequency analysis

I utilized Natural Language Processing with Python's NLTK package in order to detect which words were used most frequently in each of the four datasets individually. Word frequency analysis is a method that automatically identifies the frequency of certain words in a given text corpus. It is performed by taking the raw token from the pre-processing step and calculating the word frequencies (Menon et al., 2018).

N-grams were used to generate the cryptocurrency and used the token count function to generate the top 10 words from each coin. N-gram is a collection of word tokens from a text document so that the tokens are contiguous and occur in a sequence (Sarkar, 2019, p. 210). An n-gram is a sequence of n items from a given speech or text which gives good indicators of which words that mainly occur.

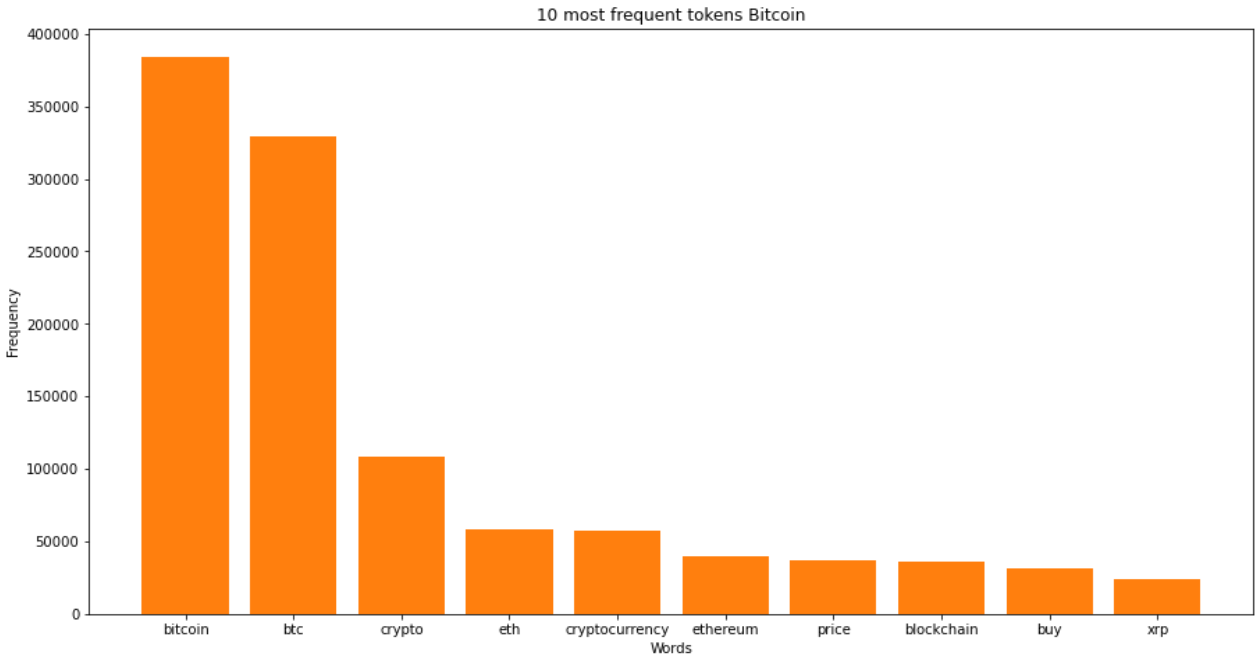


Figure 19. Overview of 10 most frequent tokens in the Bitcoin dataset

The Bitcoin dataset contains 349 654 tweets from 349 654 users where 90 066 (26%) of these users are unique. The top username in the dataset has a frequency of 4 760 (5%). The number of unique tweets are 334 074 (96%) .

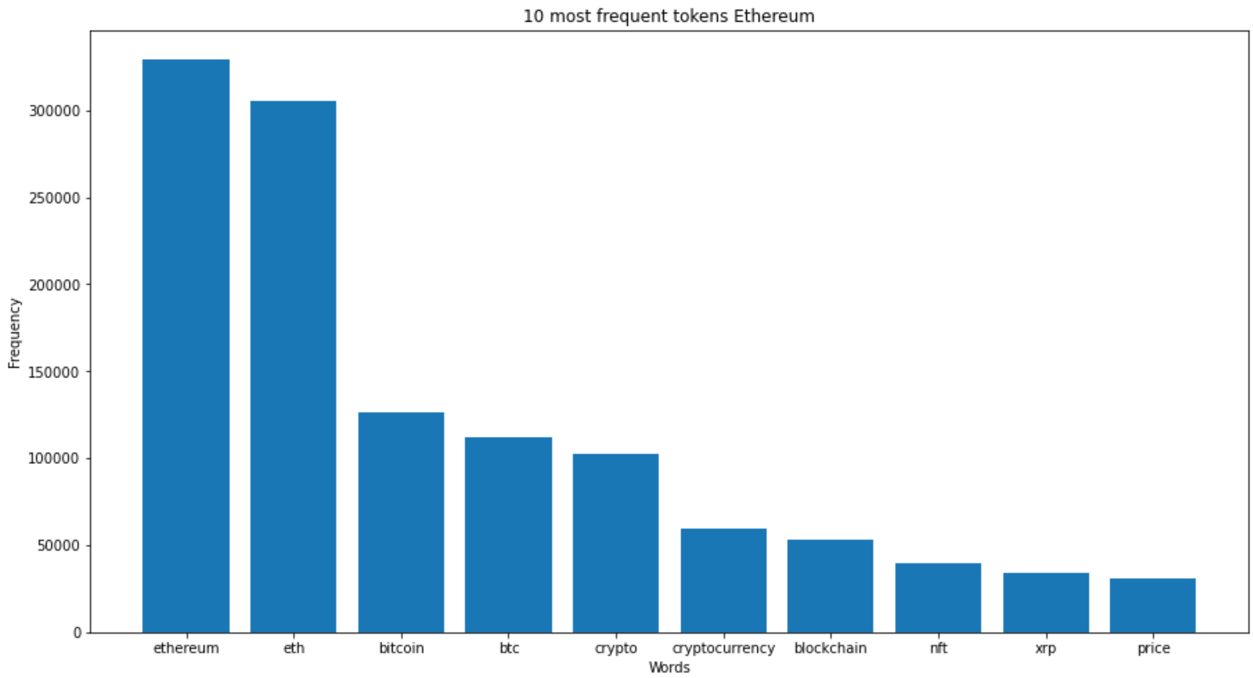


Figure 20. Overview of 10 most frequent tokens in the Ethereum dataset

In the Ethereum dataset there's 320688 tweets from 320688 users, 88585 (28%) of these users are unique. The top username in the dataset has a frequency of 3389 (4%). The number of unique tweets are 298711 (93%).

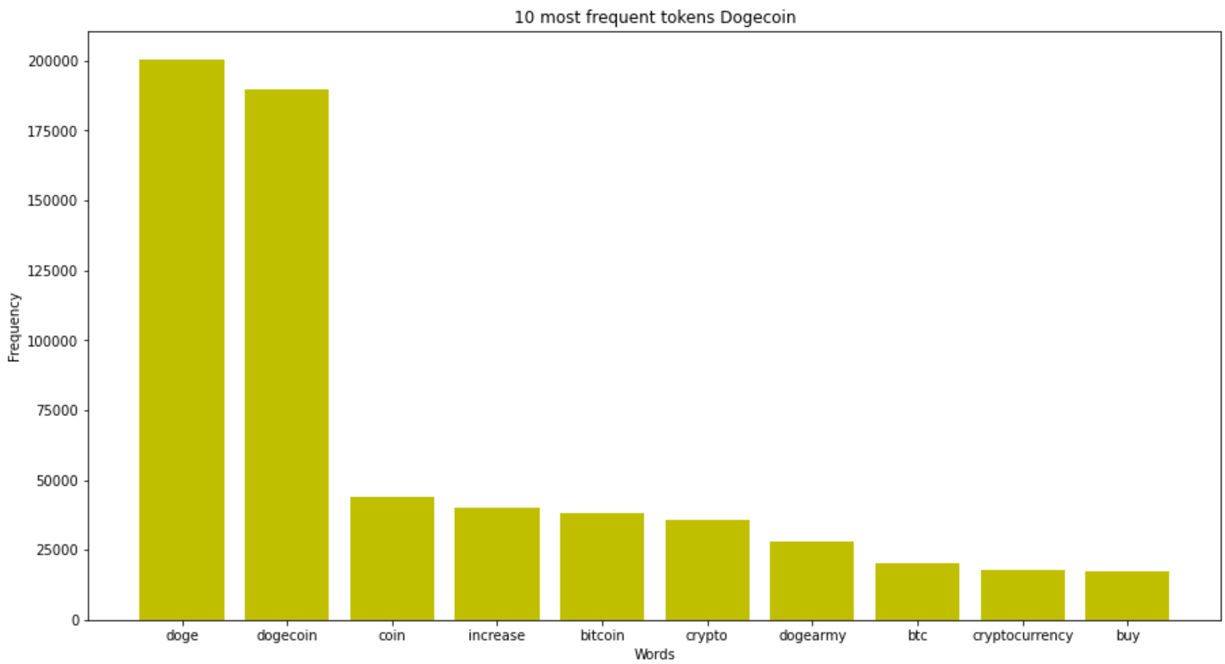


Figure 21. Overview of 10 most frequent tokens in the Dogecoin dataset

In the Dogecoin dataset with 230683 tweets from 230683 users, 86519 (38%) of these users are unique. The top username in the dataset has a frequency of 13300 (15%). The number of unique tweets are 215283 (93%).

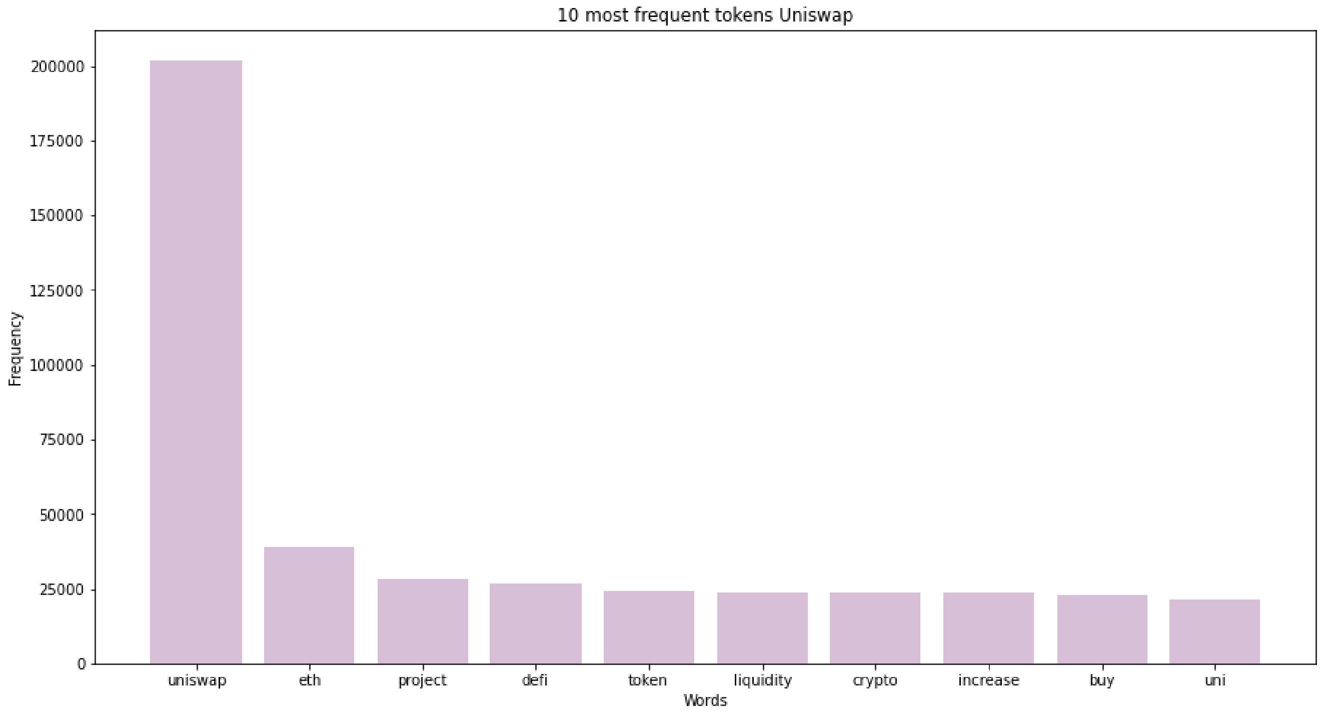


Figure 22. Overview of 10 most frequent tokens in the Uniswap dataset

The Uniswap dataset contains 291051 tweets from 291051 users, where 111579 (38%) of these users are unique. The top username in the dataset has a frequency of 1611 (1%). The number of unique tweets is 267598 (92%).

6.4 Detecting the presence of Twitter bots

Several studies and articles have mentioned the presence of bots in the context of cryptocurrencies and Twitter. In this part, I aim to explore the possibility of using a simple heuristic approach to investigate the presence of Twitter bots in the datasets. Kraaijeveld and De Smedt (2020) tested for the presence of bots by using six heuristics based on the findings in Reutzel (2018). Four of the same heuristic approaches will be implemented in this thesis. In addition, “airdrop” or airdrops” will be considered a criteria.

A tweet is considered to be posted by a bot if the text contains two or more of the following criteria:

1. “Give away” or giveaway”
2. “Register” or “join”
3. Username contains “bot”
4. Tweet containing more than 14 hashtags
5. “Airdrop” or “airdrops”

Bots that tweet “give away” or “giveaway” advertises users to transfer a small amount of money to an account in exchange for a cryptocurrency giveaway. Register or join refers to bots that ask users to register for fraudulent schemes through various links (Kraaijeveld and De Smedt, 2020). There’s a high probability that usernames containing “bot” are Twitter bots that tweet automated tweets controlled by bot software (Kraaijeveld and De Smedt, 2020). Tweets from bots often contain many hashtags, especially unrelated hashtags, which are considered spam by the Twitter platform. According to Twitter, bots often use popular or trending hashtags to manipulate a conversation or drive traffic to a website or product. Therefore, tweets containing more than 14 hashtags are a criteria (Kraaijeveld and De Smedt, 2020). An airdrop is a free cryptocurrency and token distributed through an initial coin offering (ICO) (Gao et al., 2020). Counterfeit digital tokens are also being used to carry out fake airdrops. Due to the popularity of these types of transactions, many people are becoming victims of these scams through bots (Gao et al., 2020).

In the Bitcoin dataset, 3024 usernames contain the words bot" or "bots" which indicates that the tweet was directly posted by a Twitter bot. Further, 10946 tweets met two or more of the criteria above. Thus, bots constitute 3.2% of the Bitcoin dataset. In the Ethereum dataset, 6314 usernames contain the words "bot" or "bots" which indicates that the tweet was directly posted by a Twitter bot. Furthermore, 12833 tweets met two or more of the criteria above. Thereby, bots make up 4.0% of the tweets in the Ethereum dataset. In the Uniswap dataset, 2041 usernames contain the words "bot" or "bots" which indicates that the tweet was directly posted by a Twitter bot. 6845 tweets met two or more of the criteria above. Therefore, bots constitute 2.4% of the tweets in the Uniswap dataset. In the Dogecoin dataset, 1304 usernames contain the words "bot" or "bots" which indicates that the tweet was directly posted by a Twitter bot. 9337 tweets met two or more of the criteria above. Thus, bots constitute 4.0% of the Dogecoin dataset.

Note that if a user alone has "bot" in the username, it does not fulfill the criteria in the heuristic approach. Two or more of the criteria must be met in order for a bot detection to be considered. The overview of the number of usernames containing "bot" is only included to demonstrate the number of usernames that might be directly linked.

Kraaijeveld and De Smedt, 2020 found that by applying a heuristic approach to detect the presence of cryptocurrency-related bots, an estimate of 1-14% of the tweets were bots. However, the researchers claim that the frame is an estimate and that the actual number probably is higher. The heuristic approach conducted in this paper returned an estimate of 2.4 - 4% of tweets posted by bots. This paper investigates the presence of bots through a heuristic approach, therefore only an estimated number of accounts are considered bots. Table 16 sums up the observed findings.

	Bitcoin	**Ethereum**	**Uniswap**	**Dogecoin**
Bots	3.2%	4.0%	2.4%	4.0%
# tweets meets criteria	10946	12833	6845	9337
"Bot" in username	3024	6314	2041	1304

Table 16. Percentage of detected bots in each cryptocurrency dataset

6.5 User analysis

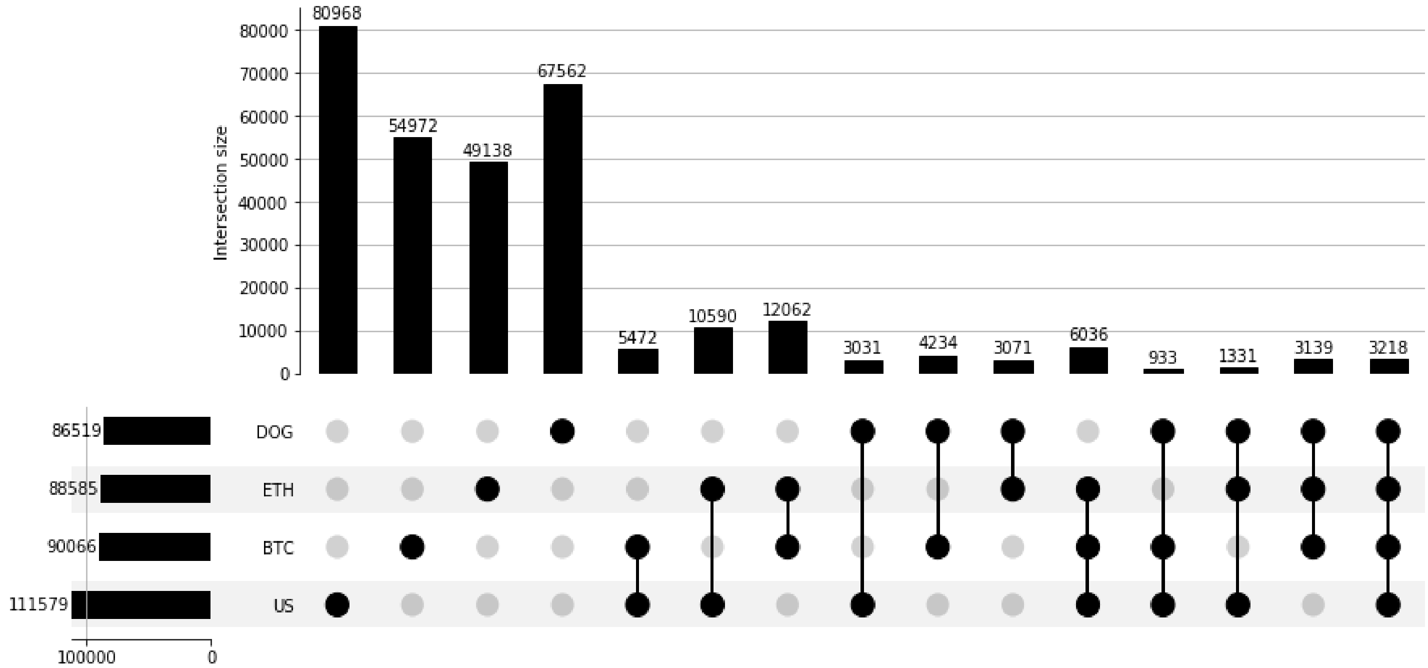


Figure 23. UpSet plot of unique users in common

The UpSet plot shows a visualization of intersecting users across the four cryptocurrency datasets. The plot shows 3218 unique users in common across all four datasets. The horizontal bars on the left indicate the total number of unique users for each cryptocurrency set. The first four vertical bar plots show the number of unique common usernames excluded from other cryptocurrencies, meaning that they are uniquely represented in one cryptocurrency. Out of the four cryptocurrencies, Bitcoin and Ethereum have the most unique users in common, with 12062 users. The smaller cryptocurrencies, Dogecoin and Uniswap, have 3031 unique users in common.

The user analysis will be conducted as follow. First, a sample of 10% will be collected from each of the four user categories uniquely represented in one cryptocurrency. Hence, the group of users that falls outside the category of common users between sets. The sample will be collected to investigate the users' sentiments and emotions and how much this group of users accounts for interactions.

Secondly, I will investigate the 3218 unique users in common across the four cryptocurrency sets. Thirdly, I will examine the unique users in common between smaller coins, Dogecoin and Uniswap, and larger coins, Bitcoin and Ethereum, to present common and distinct features between the users.

6.5.1 Sample subset

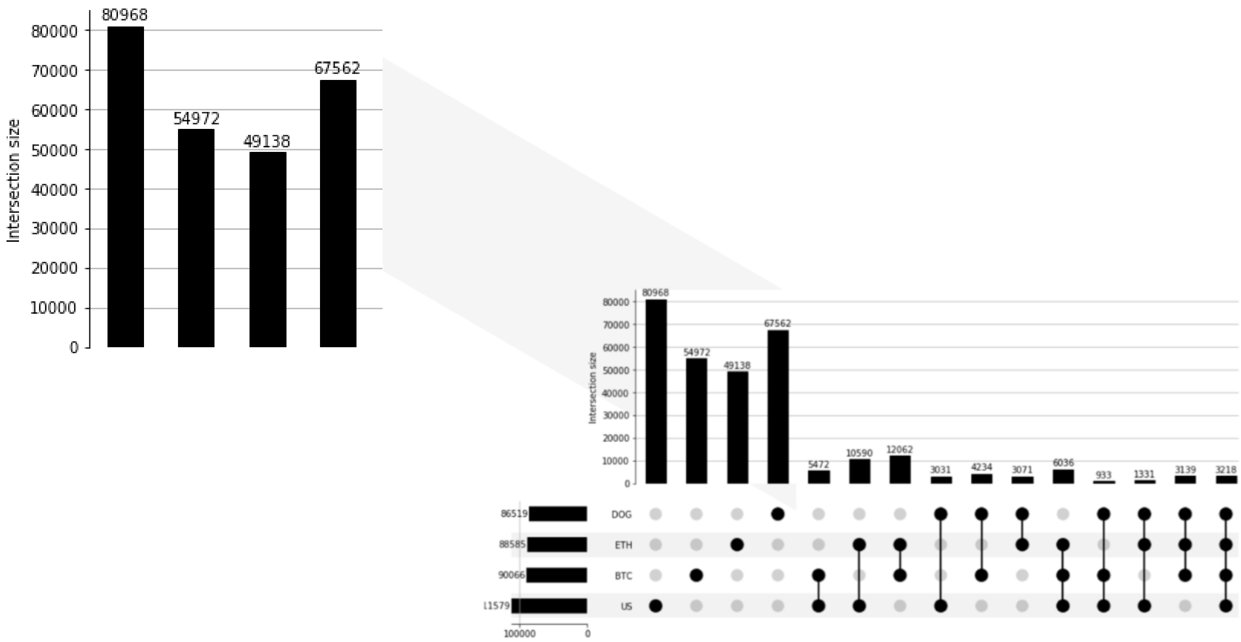


Figure 24. Highlight of the four first bar plots

A sample of 10% was collected from each set of the group of unique users that fall outside the category of common unique users between all sets, to investigate the user's interactions. Therefore, 5497 users were sampled from Bitcoin, 4914 users from Ethereum, 8097 users from Uniswap, and 6756 users from Dogecoin.

Unique users	# Tweets	# Unique tweets	Sentiments	Emotions	Frequency #5 top users
25 264	46 641	45 913	Positive 57% Neutral 30% Negative 13%	Happy 41% Fear 26% Surprise 18% Sad 11% Angry 4%	1611 319 282 268 216

Table 17. Sample data

The majority of users from the sample are labeled with positive sentiments and emotions. The group of users is labeled 57% positive and 41% happy. The sentiments of the top five users with the highest frequency of tweets in the dataset are 82% positive, and the most labeled emotion of the users is surprise 36%, followed by 35% happy. The top five users with the highest frequency of tweets in the dataset constitute 6% of the dataset.

The most discussed topics generated from the word frequency analysis show that most high-frequency words are the names of the cryptocurrencies and their ticker symbols (e.g. ‘‘BTC”). Subsequently, the cryptocurrency tickers are removed to have more topics returned. The topics discussed among the users are related to cryptocurrency, price, increase, nfts, buying and airdrop.

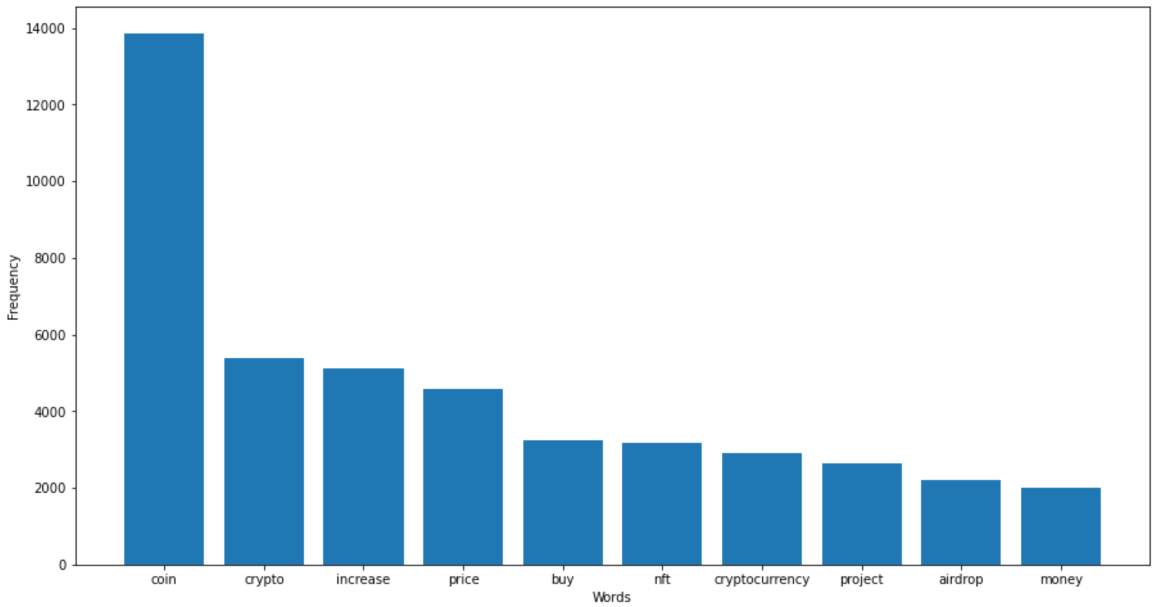


Figure 25. Top 10 frequent words after removed tickers

6.5.2 Unique users in common

# Tweets	# Unique tweets	Max tweet volume	Average tweet volume	Max word count	Average word count	Max word length	Average word length	Sentiments	Emotions	Frequency #5 top users
225 432	220 921	3049	296	51	16	254	109	Positive 44% Neutral 37% Negative 17%	Fear 36% Happy 27% Surprise 19% Sad 12% Angry 6%	7093 4704 4619 4208 4033

Table 18. Descriptive statistics of unique users in common across all four sets

The 3218 unique users in common across the four sets have tweeted a total of 225432. The top emotion amongst the users is fear, with 36%. In addition, the users are 44% positive overall. The top five users with the highest frequency of tweets are labeled 38% fear and 21% happy. The most labeled sentiment is 43% positive. The top five frequent users constitute a percentage of 11%.

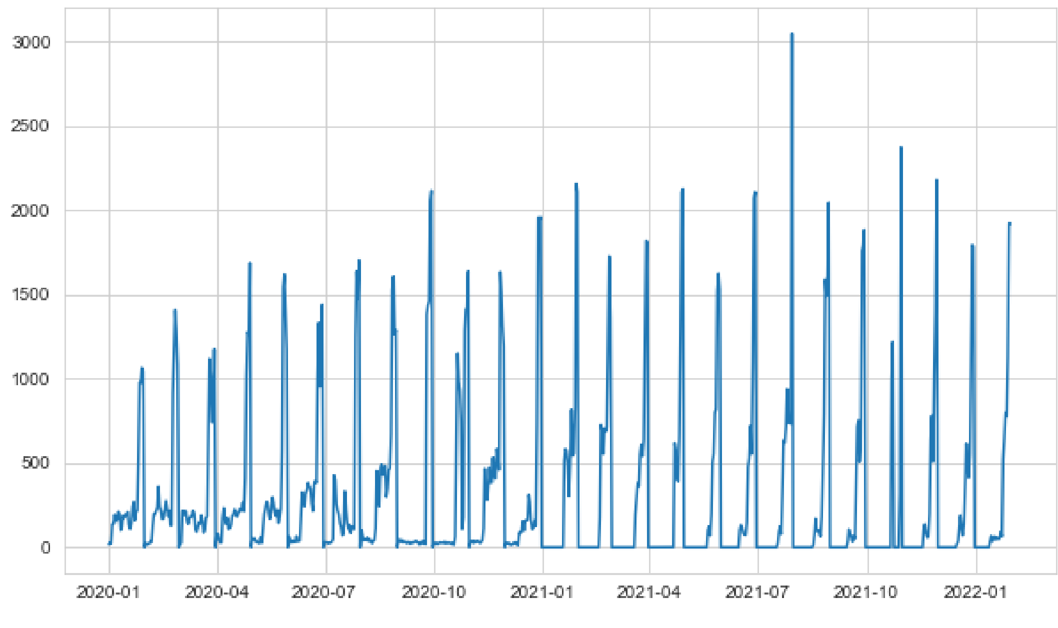


Figure 26. Tweet volume users in common

The maximum tweet volume amongst the users is 3049, while the highest tweet volumes are between July and August 2021.

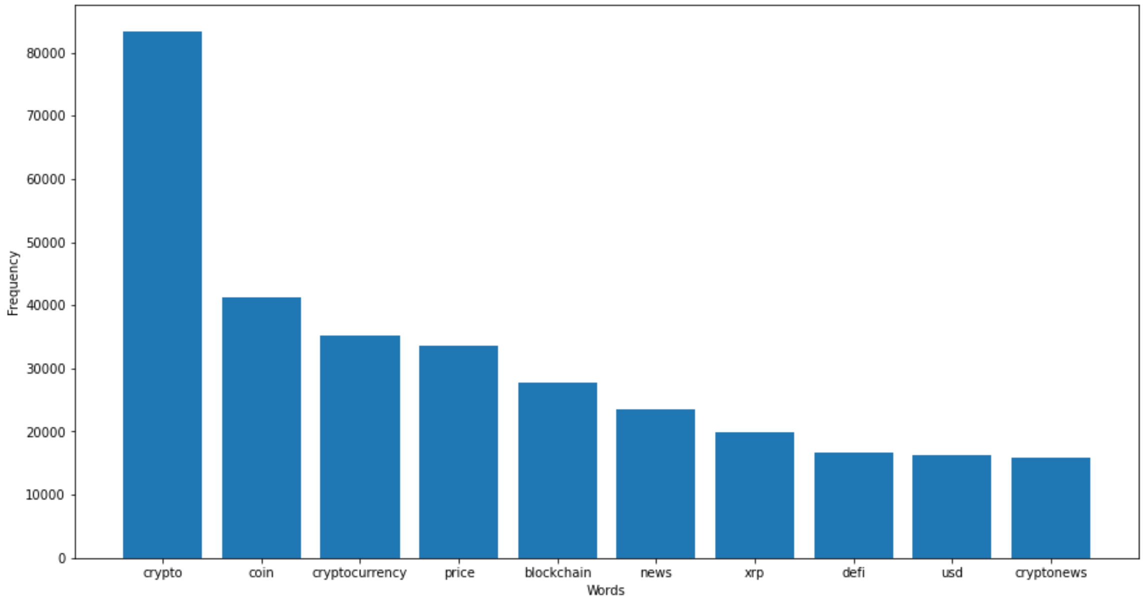


Figure 27. Top 10 frequent words

From the top 10 frequent words, the name of the cryptocurrency and tickers: "bitcoin", "btc, "ethereum", "eth", "dogecoin", "doge", "uniswap" and "uni", are excluded to get more topics in return, as the tickers have the highest word frequencies.

The most discussed topics among the unique users across the sets are primarily related to cryptocurrency news, updates, and prices with the highest frequencies in the sets. Further, a high frequency of discussions about price predictions appear among unique users across sets, 11562 times. Elon Musk has a frequency of 5386 in the tweets, where the users are discussing Musk's tweets regarding his support of Bitcoin and Dogecoin. Elon Musk is known for his technological prowess and has a large social media following. He is also known to hype cryptocurrencies and frequently uses his platform to communicate with his followers about cryptocurrencies. Per May 2022, Musk has a Twitter following of 91.4 million (Twitter.com). The users are discussing the rise in prices and reacting to tweets from Musk with "pump," which is an increase in price, or "dump" which is the price decreasing. An example of a "pump" tweet reacting to Musk's tweet is: "Musk tweets his favorite cryptocurrency pumps again!". An example of a "dump" tweet is: "Your hero Elon Musk dumped btc bitcoin".

6.5.3 Unique users Bitcoin and Ethereum

# Tweets	# Unique tweets	Max tweet volume	Average tweet volume	Max word count	Average word count	Max word length	Average word length	Sentiments	Emotions	Frequency #5 top users
90 541	86 136	2132	119	52	18	254	125	Positive 51% Neutral 32% Negative 17%	Fear 36% Happy 31% Surprise 16% Sad 10% Angry 7%	1493 756 719 709 701

Table 19. Descriptive statistics of unique users in common between Bitcoin and Ethereum

Bitcoin and Ethereum have 12062 unique users in common, which in both datasets have tweeted a total of 90541. The top sentiment among the users is 51% positive, and the most labeled emotion is 36% fear. The five users with the highest frequency of tweets are labeled 58% neutral and 45% fear. The top five frequent users constitute a percentage of 5%.

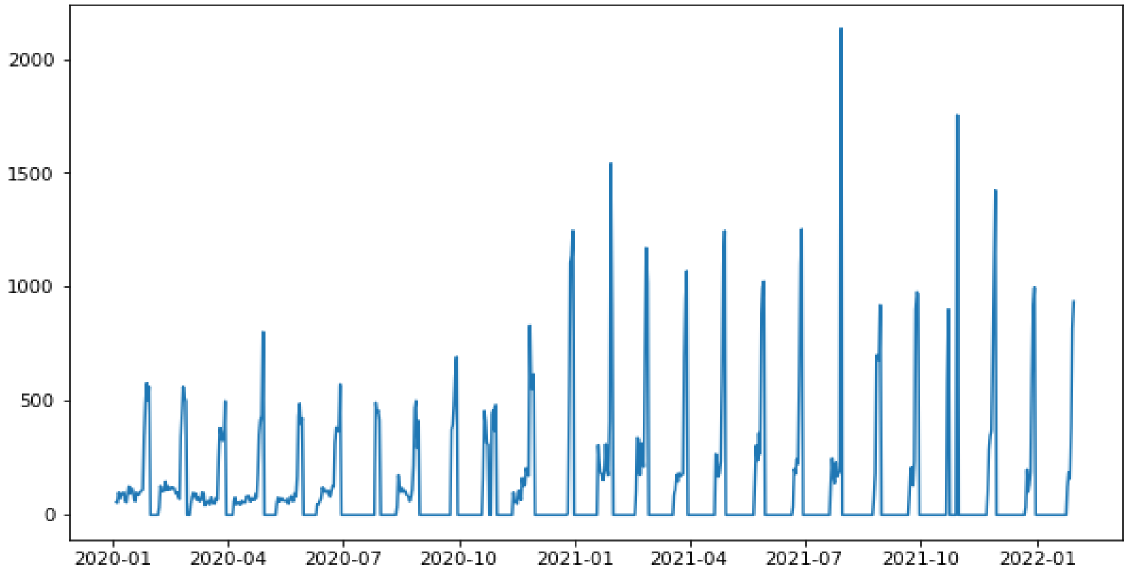


Figure 28. Tweet volume among unique users - large coins

The maximum tweet volume is 2132, where the highest tweet volumes are between July and August 2021.

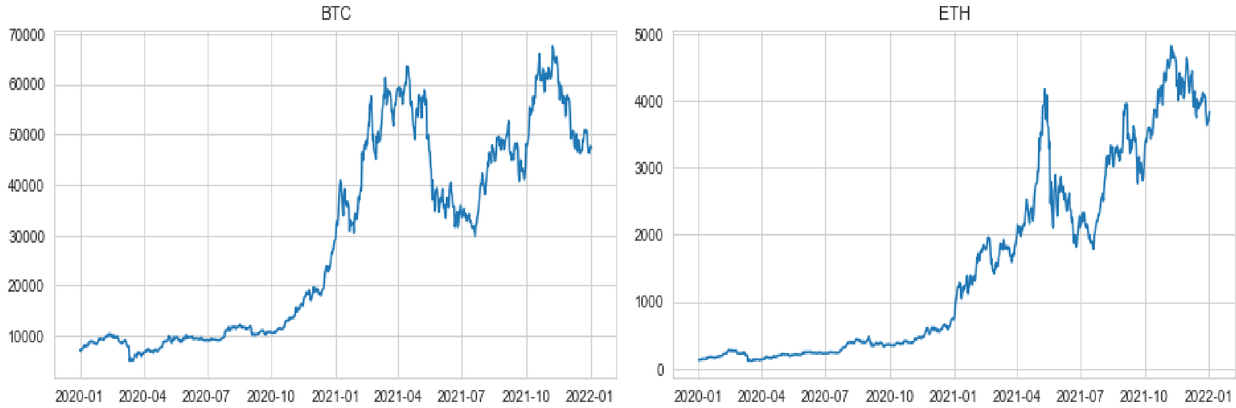


Figure 29. Time series plot BTC and ETH closing price

The time series plots show the change in cryptocurrency price in the time period of January 2020 – January 2022 for Bitcoin and Ethereum. We can observe a price dip in July 2021, where Ethereum fell below $2700 and Bitcoin below $300000 towards the end of July. Subplots are used as the coins are on different scales. The emotion amongst the users in the period of the price dip is 40% fear, and the sentiment is 52% positive. The highest tweet volume is in the same period as the price fall in July 2021.

To identify the most discussed topics amongst the users, the cryptocurrency tickers "bitcoin" (77404), "btc" (75677), "eth" (54763), and "ethereum" (50556) are excluded.

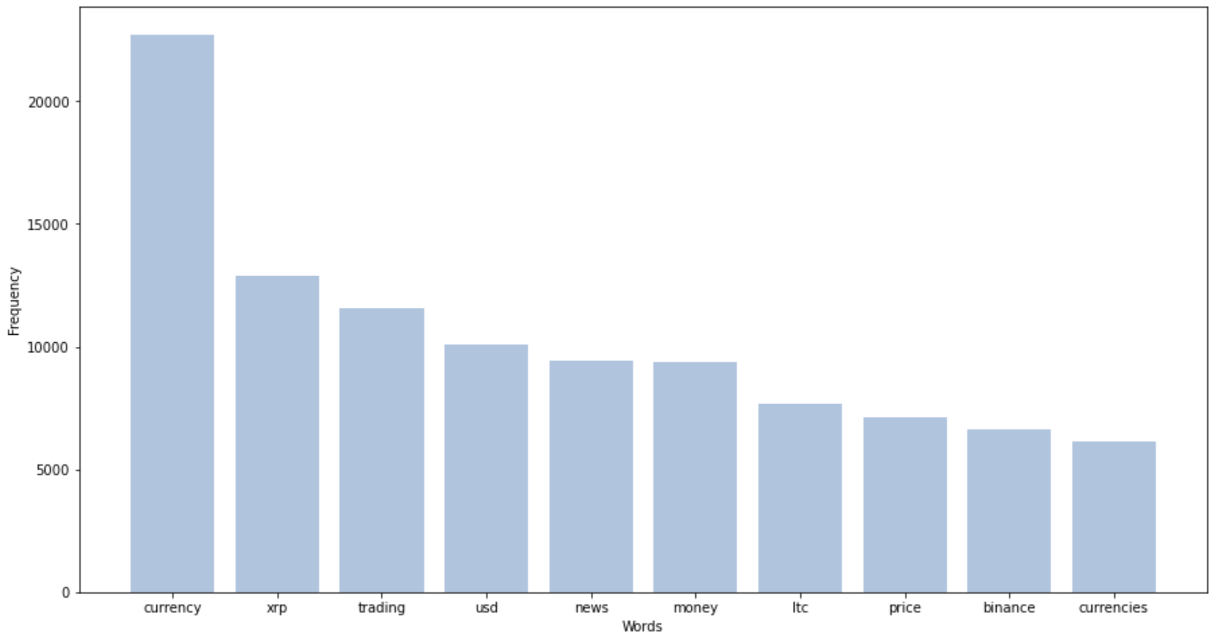


Figure 30. Top 10 frequent words among unique users - large coins

Major concepts and topics discussed among the users are related to increase, trading and price. There's a high frequency of "fomo" (1116), which refers to the fear of missing out. Since both Bitcoin and Ethereum do not pay dividends, their returns depend on increasing prices. The term "hodl" has a frequency of 1114. The term HODL refers to the idea of holding cryptocurrencies for the long term. It's also commonly used by crypto investors to describe their desire to hold on to the asset for a long time. The term "moon" also occurs in the dataset 1028 times. An investor's goal is for their assets to gain as much value as possible through reaching high valuations. When referring to cryptocurrencies, the term mooning refers to the sudden increase in the asset's value (Becker, 2021). If you look at the charts of various cryptocurrencies, it would appear that the asset's value has taken a sharp turn up during a period of increase.

6.5.4 Unique users Dogecoin and Uniswap

# Tweets	# Unique tweets	Max tweet volume	Average tweet volume	Max word count	Average word count	Max word length	Average word length	Sentiments	Emotions	Frequency #5 top users
13 901	13 830	353	18	51	13	254	83	Positive 55% Neutral 37% Negative 14%	Happy 44% Fear 22% Surprise 18% Sad 12% Angry 4%	218 168 160 92 89

Table 20. Descriptive statistics of unique users in common between Dogecoin and Uniswap

Dogecoin and Uniswap have 3031 unique users in common, which in both datasets have tweeted a total of 13901. The top sentiment among the users is 55% positive, and the most labeled emotion is 44% happy. The five users with the highest frequency of tweets are labeled 51% positive, and 48% happy. The top five frequent users constitutes a percentage of 5%.

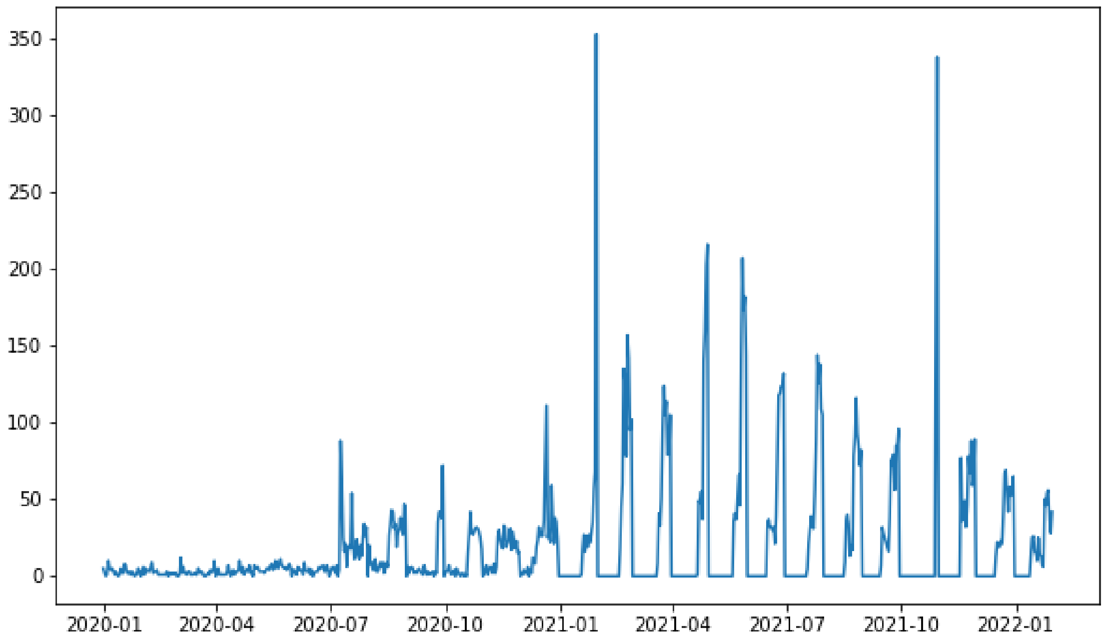


Figure 31. Tweet volume among unique users - small coins

Amongst the unique users in the Dogecoin and Uniswap sets, the maximum tweet volume is 353, where the highest tweet volume is in February 2021.

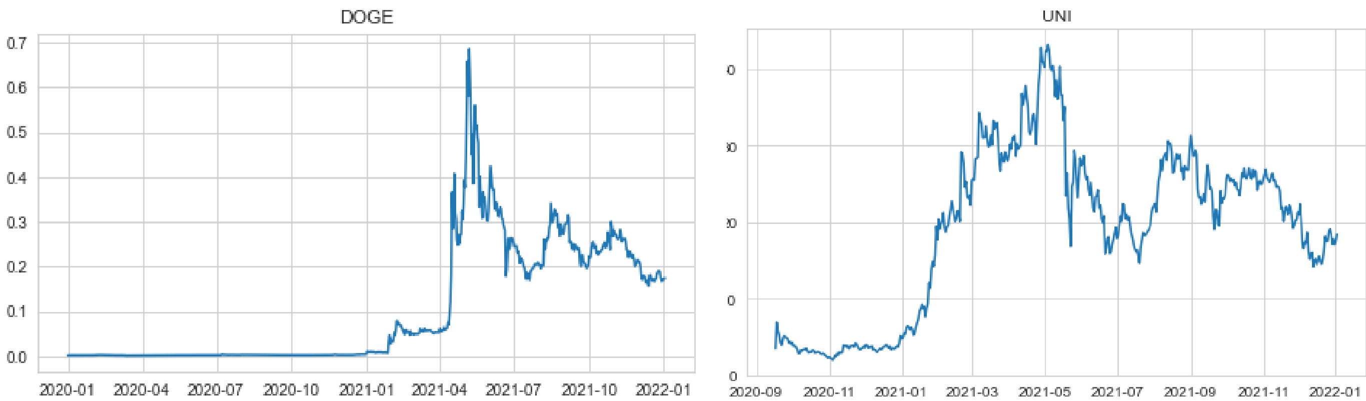


Figure 32. Time series plot Dogecoin and Uniswap closing price

The time series plots show the change in cryptocurrency price in the time period of January 2020 – January 2022 for Dogecoin and Uniswap. Subplots are used as the coins are on different scales. Uniswap prices started September 17th 2020. Dogecoin had a peak in May 2021 when the coin reached an all-time high of $0.688813 (CoinMarketCap, 2022). Uniswap peaked in April 2021and reached an all-time high with a market cap of $43.22 688813 (CoinMarketCap, 2022). Also, Dogecoin and Uniswap had a low dip in July 2021. The emotion amongst the users in this period is 49% happy, and the sentiment is 53% positive.

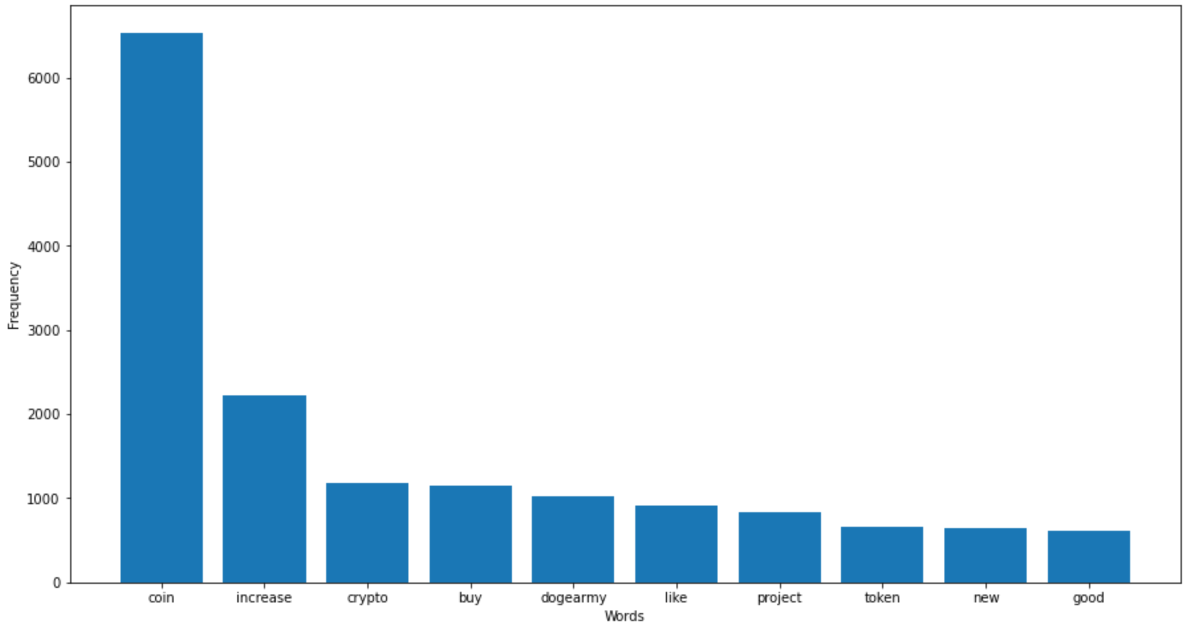


Figure 33. Top 10 frequent words among unique users - small coins

The cryptocurrency names and tickers “doge”, “dogecoin”, “uni”, and “uniswap” are excluded to get more topics in return, as the tickers have the highest word frequencies. The concepts and topics discussed amongst the unique users in common in the Dogecoin and Uniswap are associated with discussing an increase and buying of cryptocurrency. There’s a high frequency of positively charged words such as “like” and “good”. There’s also a high frequency of the word “dogearmy” which refers to fans of Dogecoin.

The discussion amongst the users is also centered around Elon Musk, “moon” and “airdrop”. An airdrop is a free distribution of an asset. Uniswap launched its token with a massive airdrop in September 2020 (Uniswap.org). Elon Musk has been called the “Godfather of Doge” because of his frequent tweets, memes and concerns regarding Dogecoin (Tandon et al., 2021).

6.6 Event study

In this part, the event study will be presented. First by defining the events of interest for each of the four cryptocurrencies. Further, the timeline of the event windows will be defined. Then the pre-events (before), events (during), and post-events (after) for each of the coins will be presented. Lastly, the identifying key findings will be presented.

6.6.1 Defining the Events

Event 1: Bitcoin

On January 28th 2021, Elon Musk, the CEO of Tesla, added #bitcoin to his Twitter profile, which immediately boosted the cryptocurrency's price (Kendall, 2022). Tesla also confirmed that it would start accepting Bitcoin as a payment method for its electric vehicles (Kendall, 2022). Shortly after, on February 8th, Tesla announced that the company had bought $1.5 billion of Bitcoin (Kendall, 2022).

Event 2: Ethereum

On November 8th, 2021, Ethereum hit an all-time high at $4.8k. A few days later, on November 11th, 2021, NFTs are taking off which generates an increase in the Ethereum price (TradingView, 2022). NFT stands for Non-fungible tokens, which means they are unique and are not interchangeable with anything else. NFTs run on smart contracts through the Ethereum network.

Event 3: Dogecoin

Elon Musk has frequently talked about Dogecoin on Twitter and referred to it as his favorite digital currency on NBC's Saturday Night Live (Browne, 2021). On December 14th, 2021, Elon Musk tweeted that Tesla merchandise would be buyable with Dogecoin. The announcement generated a price increase of 20% (Browne, 2021).

Figure 34. Elon Musk tweet (Source: Twitter.com)

Event 4: Uniswap

On September 17th, 2020, Uniswap launched its token with a massive airdrop where 15% of the total supply was distributed through an airdrop (Introducing UNI, 2020). 400 UNI[4] each were distributed to users who had used their protocol prior. An airdrop is a free distribution of an asset.

6.6.2 Defining the Event Timeline

As mentioned in the methodology, an event study considers three important time-periods or windows: the pre-event window, the event window, and the post-event window. The days of the event window are typically covered by the events that occurred during that specific period. This means that the event window can start from a certain point in time and extend into the following days (Kliger & Gurevich, 2014). Although many studies use a long event window as their starting point, the length of the window is still subject to debate. In this event study, the pre-event window is defined to be three weeks before the event window. The event window has a timeline of three weeks, while the post-event window is three weeks after the event window. There is a central point to avoid overlap. The event window is the main window of interest and should have a length where it's feasible to view the effect (Kliger & Gurevich, 2014). Therefore, the event window also has a timeline of three weeks.

[4] Uniswap tokens/coins

6.6.3 Event windows

Event 1: Bitcoin

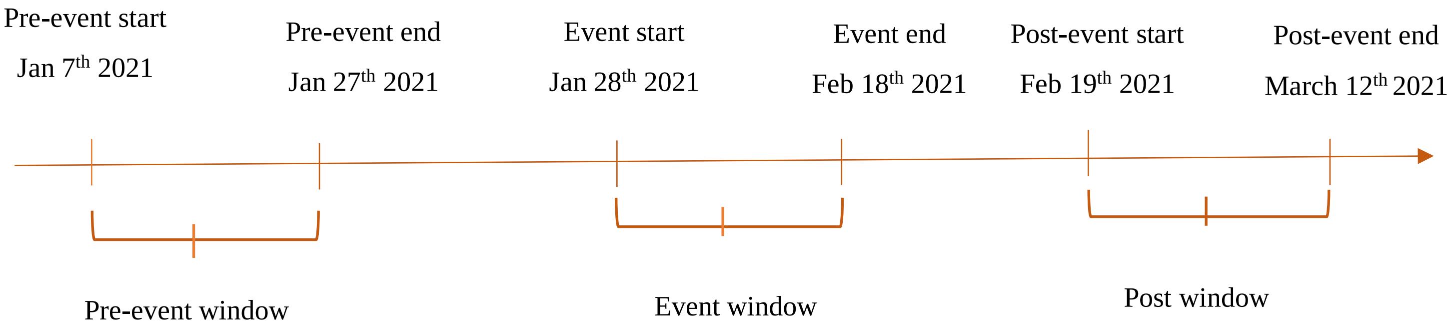


Figure 35. Event study timeline for Bitcoin. Adapted from Kliger & Gurevich (2014)

Pre-event

13685 users tweeted during the pre-event, where 6406 of these are unique users. During the pre-event, majority of the users are labeled with the emotion fear with 37%. The users' sentiment in this period is 52% positive, 29% neutral, and 19% negative.

The word frequency analysis was used to provide an overview of major concepts and topics discussed in the tweets. The topics and discussions in the pre-event are predominantly related to price discussion, users are tweeting about buying and an increase in Bitcoin. The word “price” has a frequency of 1436, “buy” (1428) and “increase”(886).

Event

The event window is the main period of interest in the event study. 19985 users tweeted during the event, where 9795 of these are unique users. 38% of the users were labeled happy. The users' sentiment during the event is 51 % positive, 32 % neutral, and 17 % negative.

From the observed topics and discussions generated from the word frequency analysis, there is a high frequency of the words “buy” (3743), “increase” (2588), “price” (2337), and “news”

(2299). The discussion about a price increase heightened from the pre-event window to the event window.

1856 tweets contained “Elon” or “Musk” during the event. Out of the 1856 tweets about Elon Musk in the event period, 48% of these are labeled positive. The most labeled emotion in the Elon Musk discussion is happy, followed by surprise. “Twitter bio” has a frequency of 247 tweets where the discussion is directly related to the fact that Elon Musk changed his Twitter bio to #bitcoin. In addition, Tesla and SpaceX have a frequency of 620. From the pre-event timeline to during the event, the emotions have shifted from fear to happy. The sentiments are still around 50% positive. Thus, the percentage of positive tweets remains approximately unchanged, and negative sentiments have decreased slightly.

Post event

14026 users tweeted during the post event, where 6764 of these were unique users. During the post event, 36 % of the users have the emotion of fear. However, the overall sentiment of the users is 50% positive, 34 % neutral, and 16% negative.

There are fewer words containing “Elon” or “Musk” in the post event. There’s now a frequency of 432, where the top sentiment amongst users is labeled 48% positive. In the post event, the top topics are centered around an increase of Bitcoin with a frequency of 1371, followed by discussions about trading, money and prices. In addition, there are discussions around altcoins such as Ethereum, Dogecoin, XRP, Ada and Litecoin. There are discussions around decreasing and lying with a frequency of 894. This indicates that the discussion during the event has continued in the post event, thus with a shift. From the timeline of the event to the post-event, the most expressed emotion has shifted from happy to fear.

Users across time

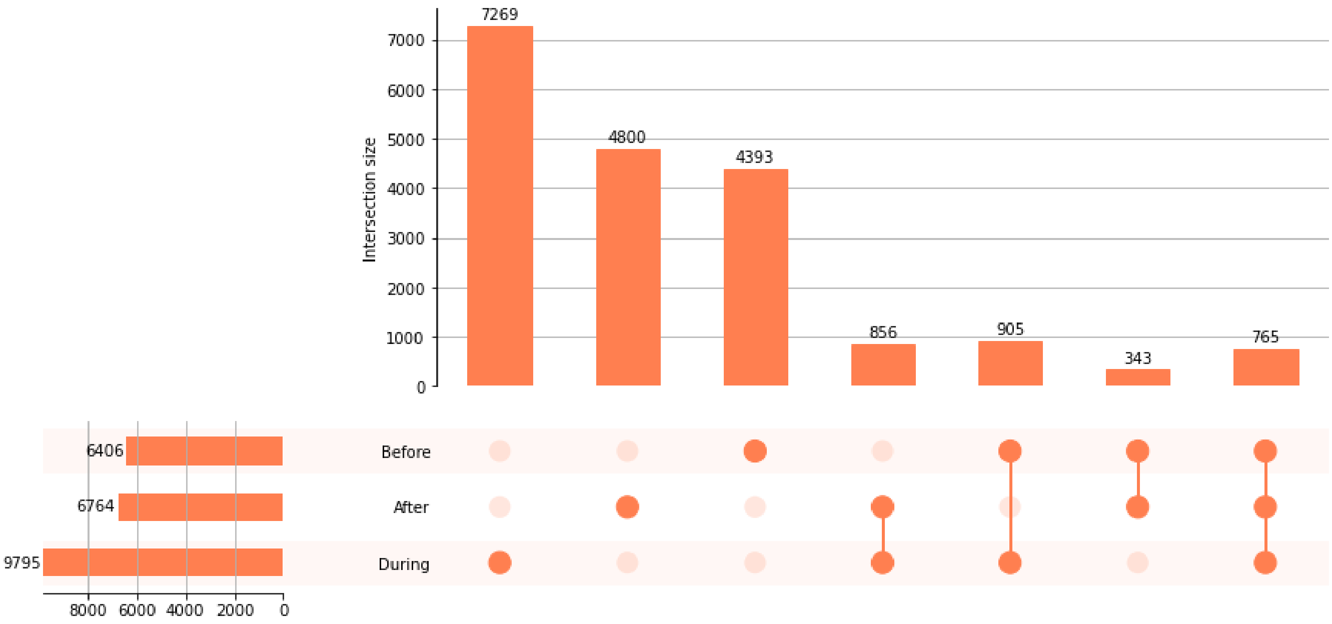


Figure 36. Unique users in common Bitcoin event

Figure 36 describes the number of unique and common users before, during and after the event. There's a total of 765 unique users in common across the three windows. There's a larger interaction of users during the event window.

Event 2: Ethereum

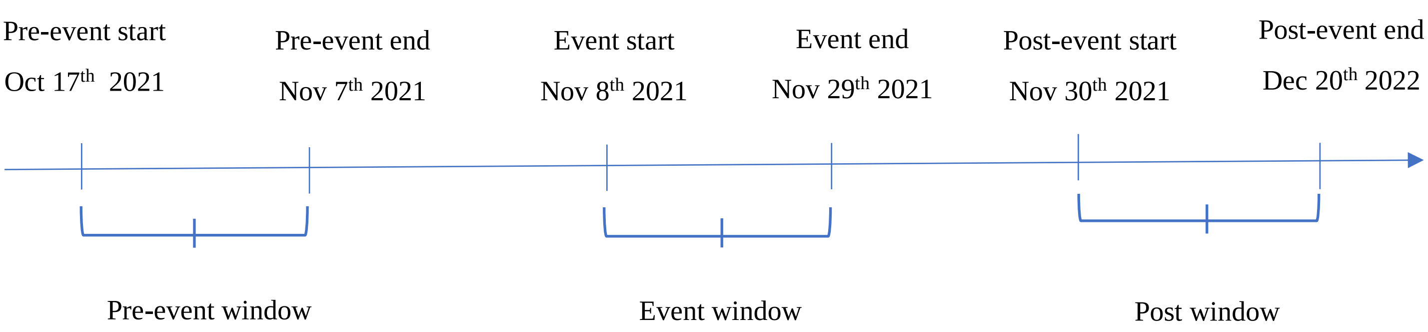


Figure 37. Event study timeline for Ethereum. Adapted from Kliger & Gurevich (2014)

Pre-event

15728 users tweeted during the pre-event with a total of 10125 unique users. During the pre-event, most users are labeled with the emotion happy 35% followed by fear 33%. The users' sentiments in this period are 49% positive, 38% neutral and 13% negative.

The most discussed topics in the pre-event are NFTs with a frequency of 4087, also “art” (2949), “nftcommunity” (1099) and “digitalasset”(1015) has high frequencies. The sentiments towards the discussion on NFTs are 44% positive. The most labeled emotion in the discussion is 43% happy, followed by 31% fear. The discussion is also centered around cryptocurrency price, wish for an increase and DeFi (decentralized finance).

Event

During the event, 14702 users tweeted. 7607 of these users are unique. During the event, most users are 56% positive, 29% neutral, and 15% negative. 37% of the users are happy, followed by fear, 28%, and surprised by 19%.

The most discussed topics during the event window are related to NFTs (4102), Opensea (1582), and gas fees (1262). Opensea is the largest NFT marketplace. The sentiments towards the discussion are 53% positive, and most emotions are labeled happy, 54%, and fear 32%. A gas fee is a cost for transactions on the Ethereum blockchain. The gas fee increases as the network get more complex due to the amount of transactions it handles (Ethereum.org). Fear is the most labeled emotion in the discussion of gas fees.

Further, the discussion is related to the increase in price for Ethereum, with the frequency of words such as “price” (1375), “increase” (1261), and “defi” (1254). The sentiments towards the discussion concerning the increase of the Ethereum price are 56% positive. The majority of emotions are labeled happy, 38%, followed by fear, 30%. From the pre-event timeline to the event, the positive sentiment has increased by 7%.

Post event

There are 11851 users in the post event, where 5909 of these are unique users. During the post event window, most sentiments are labeled positive, 47%, followed by neutral at 35%, and negative sentiments at 18%. The most labeled emotions among users are 38% fear and 39% happy.

The topics discussed by the users still revolve around NFTs and the nftcommunity, with a frequency of 3472. The sentiments towards the NFT discussion are now at 42% positive. Users seem to be more interested in price discussions, with a frequency of 2471. In the post event, there is a price discussion about Ethereum decreasing in price, with a frequency of 668. The sentiments in these tweets are predominantly negative, and the emotions are labeled with fear. From the event's timeline to the post-event, the sentiment is still positive but has decreased slightly. The emotion has shifted to 38% fear, while users are being labeled happy in both the pre-event and the event.

Users across time

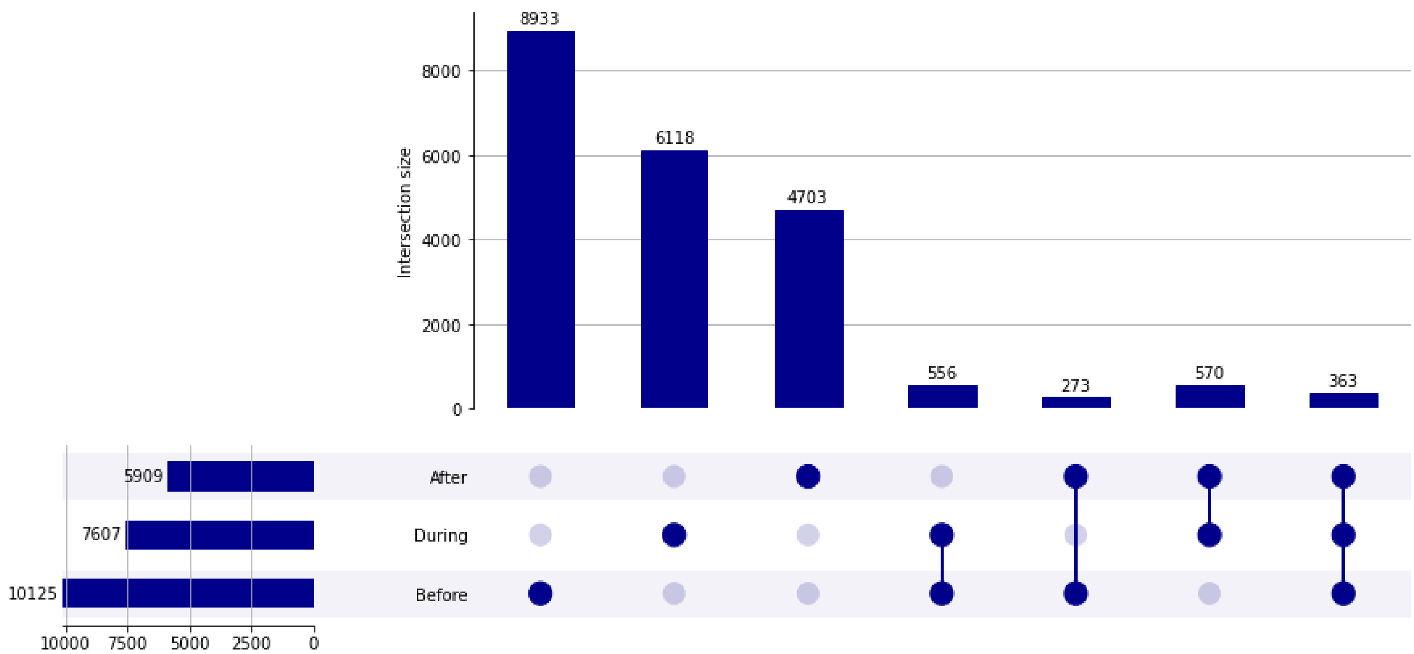


Figure 38. Unique users in common Ethereum

There are 363 unique common users involved in all the three stages of the event. The before window has a larger interaction than during and after the event. Thus, there are more unique users in common during and after.

Event 2: Dogecoin

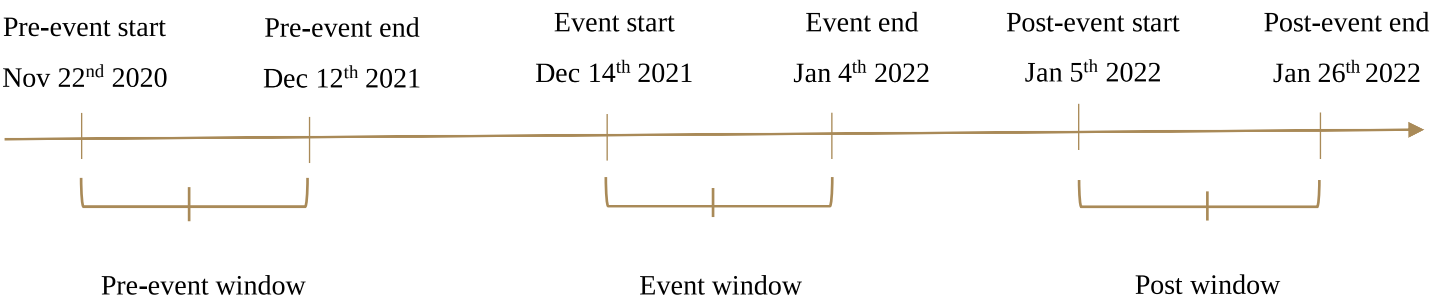


Figure 39. Event study timeline for Dogecoin. Adapted from Kliger & Gurevich (2014)

Pre event

11050 users tweeted during the pre-event, where 5560 of these were unique users. The majority of the users are positive 59%, neutral, 28%, and negative 13%. The top emotion is happy 36%, followed by sad 22%.

The most discussed topics in the pre-event are related to a wish for an increase in Dogecoin. The word “increase” has a frequency of 2521. There are discussions about other meme coins, such as Shiba Uni, with a frequency of 1387. The sentiment in the discussions about Shiba Uni is 55% positive. Elon Musk appears in the discussion 562 times in the pre-event. The sentiment in the discussion about Elon Musk is 43% positive, and the most expressed emotions are 44% happy and 23% surprise.

Event

During the event, there were 8463 tweets where 3809 of the users were unique. The most common sentiment is positive at 60%, followed by 30% neutral and 10% negative. The most labeled emotion is happy, 40%.

The discussion during the event is centered around the increase of Dogecoin. 1658 tweets contain “moon”, which often is used to describe a cryptocurrency's rise in value. Another common use of the phrase is “to the moon,” which suggests that an asset will rise in price soon (Mitra, 2021). Another topic is the increase of Dogecoin, which has a frequency of 1231 during the event. Elon Musk appeared in the discussion 878 times during the event. 57% of the discussions in regard to Elon Musk are positive. The most labeled emotion in this discussion is happy 29%, followed by surprise 27%.

Post event

In the post event, there are 11012 users, 4414 of these users are unique. During the post event, the majority of the users are labeled with the emotion happy 41%, followed by surprise at 21%. The users' sentiment in this period is 53% positive, 30% neutral, and 17% negative.

In the post event, the topic of Tesla and Elon Musk continues. The words with the highest frequencies are “increase” (1353), “tesla” (1198), and “accept” (10445). The discussion around “accept” is linked to the Elon Musk tweet in the event window about accepting Dogecoin as payment for Tesla merch. The sentiments in the discussion are predominantly positive.

However, the conversation topic shifts in the post event. On January 25, 2022, Elon Musk posted a tweet that said he would eat a happy meal on TV if McDonald's accepted Dogecoin. Therefore, “McDonald's” also appears in the post event with a frequency of 593. The sentiments in the discussion are 77% positive, and the emotion among users is 43% happy. The words “accept” and “increase” are also in the context of this event. From the timeline of the event to the post-event, the positive has sunk by 7%. The percentage of negative users increased slightly after the event.

Users across time

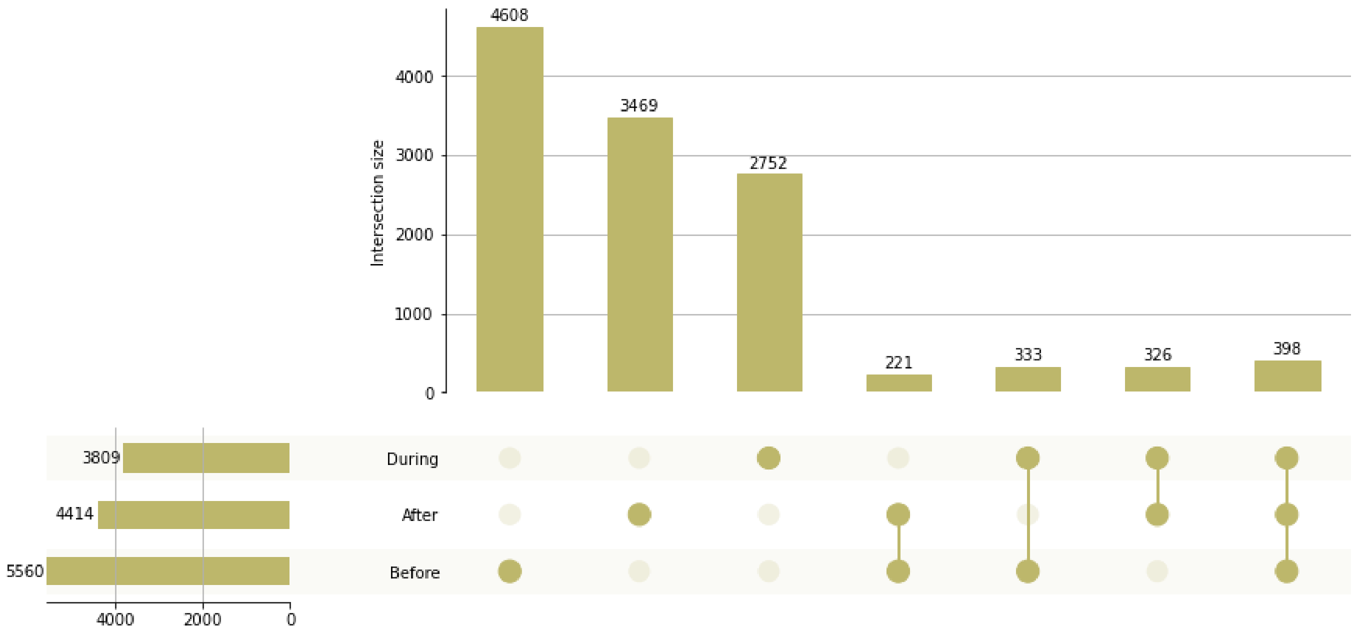


Figure 40. Unique users in common Dogecoin

There’s a total of 398 unique users in common across the three windows. There’s a larger user interaction before the event than during and after. There are more users in common during and after the event.

Event 4: Uniswap

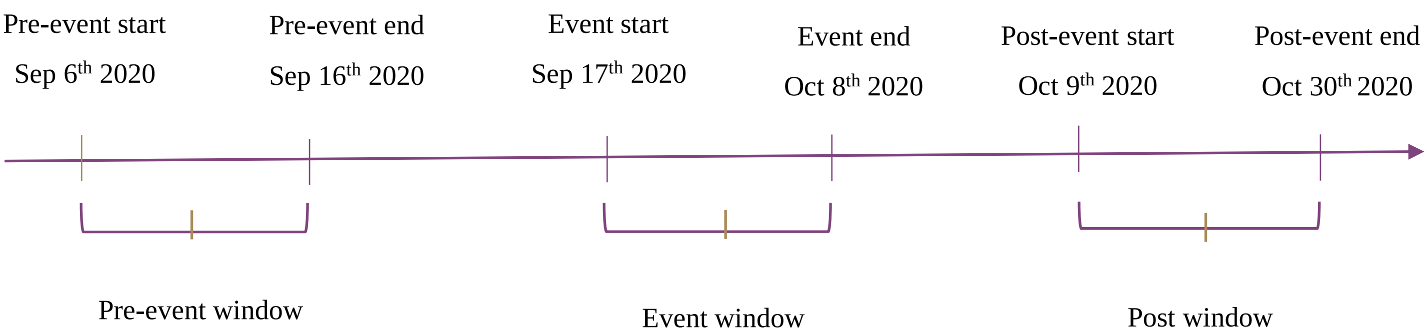


Figure 41. Event study timeline for Uniswap. Adapted from Kliger & Gurevich (2014)

Pre-event

In the pre-event, there were 12758 users and 6358 of these were unique. Most users are labeled with the emotion happy 32%, followed by fear 30%. The users' sentiments during the pre-event are positive 52%, neutral 30%, and negative 18%.

The topics discussed in the pre-event are related to liquidity, price, and trading. In addition, there's a high frequency of words such as "defi" (2273), "liquidity" (1370), and "buy" (1145). There's also a frequency of the word "airdrop" (992) where users are anticipating an airdrop. The sentiments in this discussion of an airdrop are 84% positive.

Event

During the event, 12524 users tweeted, 4346 of these were unique users. The majority of the users are positive 69%, neutral, 19%, and negative 13%. The top emotion amongst the users is surprise 34%, followed by happy 31%.

The most discussed topic during the event was the airdrop, with a frequency of 2635. The sentiment related to the discussion of airdrops is 89% positive, and the top emotion amongst the users is happy 79%. Further, the most common words during the event are "thanks" and "thank" with a total frequency of 3761. In addition, the word "amazing" occurs 1171 times. This indicates that the users are showing gratitude for the distributed airdrop. From the pre-event timeline to the event, the positive sentiment has increased by 17%.

Post event

12361 users tweeted in the post event. 7541 of these users are unique. During the post event, 57% of the users are positive, 27% are neutral, and 17% are negative. The most expressed emotion amongst the users are happy 41% and surprise 26%.

In the post event, the topic of discussion is a continuation of the event where the massive airdrop is still discussed, now with a frequency of 1736. The word with the highest frequency is "thank"

2672. The sentiments amongst the users in the discussion of the airdrop are 68% positive, and 47% of the users are happy, which is the most labeled emotion in the discussion. In addition, there’s a price discussion amongst the users where “liquidity”, “increase”, “price” and “buy” have high frequencies. The price discussion is 69% positive, indicating that the users are hopeful for a price increase.

Users across time

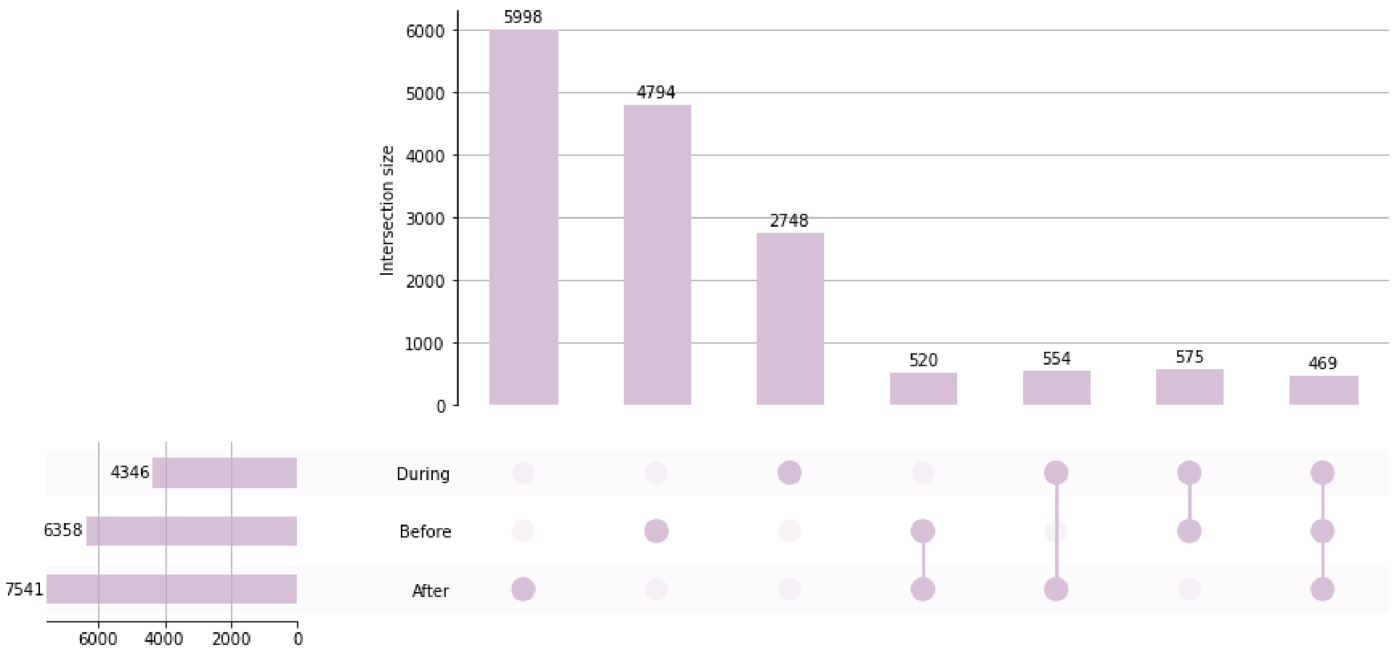


Figure 42. Unique users in common Uniswap

There are 469 unique common users involved in all the three stages of the event. The after event window has a larger interaction than before and during the event. There’s more users in common before and during the event.

7 Discussion

In this section the findings conducted in the analyses of the multiple-case study will be presented and discussed. This section aims to answer the research questions based on the findings in the conducted data analyses.

7.1 Different social media user interactions around cryptocurrencies

Statistics	Bitcoin	Ethereum	Uniswap	Dogecoin
# Tweets	349 654	320 688	291 051	230 683
# Unique tweets	334 074	298 711	267 598	215 283
# Unique users	90 066	88 585	111 579	86 519
Bots	3.2%	4.0%	2.4%	4.0%
Sentiments	Positive 49 % Neutral 33% Negative 18%	Positive 51% Neutral 34% Negative 15%	Positive 57% Neutral 29% Negative 14%	Positive 55% Neutral 32% Negative 13%
Emotions	Fear 35% Happy 33% Surprise 16% Sad 12% Angry 4%	Fear 35% Happy 31% Surprise 17% Sad 11% Angry 6%	Happy 37% Fear 28% Surprise 20% Sad 11% Angry 4%	Happy 44% Fear 19% Surprise 17% Sad 16% Angry 4%

Table 21. High-level analysis

The high-level analysis provides a reflection of the findings regarding the user interactions on the different cryptocurrencies. There is a deviation between the number of tweets and the number of unique tweets in the datasets. The deviation is 4% at the lowest and 8% at the highest. The deviation may be due to the occurrence of Twitter bots in the dataset.

As a result of the sentiment analysis, both VADER and TextBlob returned the highest level of positive sentiments across all the four cryptocurrencies. Additionally, the level of neutral tweets is also higher than negative. The positive sentiments are between 49% at the lowest, and 55%

at the highest among all the cryptocurrencies. In the study by Abraham et al. (2018), the findings revealed that when sentiment analysis is conducted, it tends to find tweets to be more neutral than the public sentiment. This makes the results less efficient as the sentiment does not indicate a clear pattern. The study concluded that Twitter sentiments regarding cryptocurrency tend to be positive despite price changes (Abraham et al., 2018). Also, the study conducted by Narman and Ulu (2020) observed more positive sentiments than negative sentiments in their research on cryptocurrencies. However, the degree of positivity and negativity varies depending on the currencies. The sentiments in all datasets are generally a majority of positive sentiments. A study by Kennedy et al. (2006) found that lexicon-based methods generally tend to have a positive bias. The smaller coins, Dogecoin and Uniswap are higher in positive sentiments and the most expressed emotion is happy. The larger coins, Bitcoin and Ethereum are lower in positive sentiments compared to the smaller coins, in addition the most expressed emotion is fear.

The findings in the word frequency analysis reveal that the most high-frequency words have a close relationship to their respective categories, hence other cryptocurrencies. Amongst all the cryptocurrencies, the token name and ticker (e.g. ‘‘btc”) has the highest frequency. There are high frequencies of other cryptocurrencies in the overviews of the 10 most frequent words in each dataset. The findings also show that the same topics are recurring in the different cryptocurrencies. The word frequency analysis returns a high frequency of either the word “price”, “buy”, or both in all four cryptocurrency datasets. This indicates an ongoing discussion of buying tokens and the prices of the currencies. Researchers have studied the effect of word-of-mouth. For instance, Phillips and Gorse (2022) state that word-of-mouth is believed to be an important factor in making an investment decision. It has been shown that investors are more likely to share their characteristics with their peers. In the digital age, this can be linked to online discussions on social media and may explain the recurrence of the same topics across all coins.

In the Bitcoin dataset, “eth” is the fourth and “ethereum” the sixth most mentioned word. Mentions about Ethereum have a total frequency of 97216 (28%). In Ethereum, “bitcoin” is the second most mentioned word and “btc” the third. Bitcoin has a total frequency of 238688 (74%) in the dataset. In the Dogecoin dataset, “bitcoin” is the fifth and “btc” the eighth token with the highest frequency. Bitcoin has a total frequency of 58021 (25%). In the Uniswap dataset, Ethereum is the second word with the highest frequency and appears 38812 times in the dataset

(13%). There is a connection between the cryptocurrencies, and there is a similarity where the word frequencies refer to the larger cryptocurrencies Bitcoin and Ethereum. Bitcoin is the most mentioned cryptocurrency across the sets. Bitcoin is often viewed as the driver of other cryptocurrencies. A research by Kumar and Anandarao (2019) showed that Bitcoin has significant spillover to other cryptocurrencies, for instance Ethereum.

7.2 User interaction in larger coins compared to smaller coins

Dogecoin and Uniswap are significantly smaller than Bitcoin and Ethereum in market cap, making them smaller cryptocurrencies. Dogecoin was created as a meme coin and joke of Bitcoin, Uniswap was created to enable users to trade cryptocurrencies without third-party involvement. Bitcoin was created as a currency to store and manage digital assets, and Ethereum was created as a decentralized network to distribute processing power.

The findings in the user analysis demonstrate the analytical utility of UpSet and Social set analysis. The findings show more unique users in the Ethereum and Bitcoin sets. The users have a higher tweet frequency and volume than the unique users in common in the Dogecoin and Uniswap sets. In addition, the users in Dogecoin and Uniswap have a shorter word count and word length in the tweets compared to the users in the Ethereum and Bitcoin sets. This illustrates that the users generally have shorter tweets in the discussion of smaller coins. The most labeled sentiments across the users in both small and larger coins are mainly positive with just over 50%, while the emotions differ. Unique users in common between the smaller coins are mostly labeled with 44% happy, followed by 22% fear. While the emotions amongst the users tweeting about the larger coins are labeled as 36% fear, followed by 31% happy. The smaller coins are generally more positive in both sentiments and emotions. The findings of Aslam et al. (2022) revealed that the majority of the Twitter users had the emotions happy, followed by fear and surprise in regards to the study on cryptocurrency-related tweets.

All of the four cryptocurrencies had a price fall in July 2021. Thus, the users have different sentiments and emotions regarding this. The unique users in common in the Ethereum and Bitcoin sets are predominantly labeled with 40% fear in the period of the price dip, however the sentiment is 52% positive. The unique users in common in the Dogecoin and Uniswap sets are 49% happy in the same period, and the sentiment among the users is labeled 53% positive.

This indicates that the users in the smaller coins showed a dominant positive sentiment during the period, and were more hopeful and positive despite a price dip. In a study by Burggraf et al. (2020), the results show that when investors are more pessimistic, Bitcoin declines. This is similar to the behavioral finance theory of Demiret al. (2018), which states that the relationship between cryptocurrencies and uncertainty is negative. Moreover, this can be linked to investors' risk-taking behavior concerning larger coins. The result of the user analysis found that this theory relates to larger coins to a greater extent than to smaller coins.

Dogecoin was initially created of memes in social media. Meme coins go hand in hand with social media because it has its establishment and growth directly from social media platforms. The CEO of Binance, the world's largest cryptocurrency exchange platform, stated that due to its relatively small size, Dogecoin is an ideal target for social media users to reach a larger audience that can significantly impact the price (Zhao, 2021). Because of the popularity of Dogecoin and Uniswap on social media platforms, their price tends to spike with a spike in conversation. Therefore, smaller coins are more prone to market sentiment (White-Gomez, 2022).

The topics among the unique users in common are related to cryptocurrency news, updates, and prices. At the same time, unique users among the larger cryptocurrencies discuss topics related to an increase, trading and price, which indicates that the discussion is related to the cryptocurrency market and financial assets. A recent publication by Ortu et al. (2022) studied cryptocurrencies in the environment of social media, where the influence of the online discussion of the two largest cryptocurrencies Bitcoin and Ethereum are considered. The study found the recurrence of topics and sentiments in both cryptocurrencies and that the influence of the user-generated content effected the cryptocurrency market.

Among the unique users in the smaller coins, the main topics are relates to a price increase and investments, but with a higher frequency of positively charged words. Based on the findings, one can assume that sentiments and emotions towards smaller coins are more positive because conversations and discussion topics among smaller coins seem to build more excitement and hype through tweets, which seem to be contagious among the users. For example in Dogecoin, the users engage in conversation with one another by tweeting "dogearmy", which refers to a group of Dogecoin fans on social media who are known for hyping the coin. Smaller coins seem to have a higher degree of hype-based discussions. Sociologists believe that the failure and success of cultural markets can be attributed to the various social feedback processes that occur during the hype, these include the activation of fear or excitement (Salganik et al., 2008). In the case of cryptocurrency, Jahani et al. (2018) describe this as a hype-based social process where smaller coins gain popularity due to the excitement generated by the community.

7.3 Important events effect on the user interactions

The event analysis investigated three time-periods of before, during and after in relation to important cryptocurrency-related events. The findings in the event study present three key results. The first key result is that the findings show a temporal dynamic shift regarding sentiments and emotions, where events indeed affect user interactions. The event study shows a consistent shift, where an event affects the conversation topics, sentiments, and emotions compared to the before and after window. The comparison between the pre-events, events, and post events shows that the timeline before the discussion and topics return to the state of the pre-event is a short window. For example, in the Bitcoin event, the emotions shifted from 37% fear in the pre-event, to 38% happy during the event to 36 % fear in the pre-event window. The same pattern is recurring regarding sentiments in all three windows, where the sentiments shift during the event but return to approximately the same percentage in the post-event as in the pre-event. The change is therefore temporal. For example, in the Uniswap event, 52% of the users were positive in the pre-event, 69% positive during the event, and 57% positive during the post-event. The results show that approximately 50% of the tweets are positive in the pre-event, and that there are more neutral tweets than negative. Therefore, this may indicate that the changes in sentiment and emotions require a timeframe. Thus, there exists a relationship between events and attitude. However, the relations vary according to the currency. Case in point, a study by Mukkamala et al. (2017) found that temporal distributions of user interaction in social media are a structural indicator of net positive.

The second key result is that the representation of a topic changes throughout the three-event windows. The evolution of topics throughout the windows is viewed in the context of word frequencies. The frequency of trending topics occurs during the pre-event, peaks during the event, and decreases somewhat in the post event. For example, during the Uniswap event, the topic of the airdrop was already anticipated in the pre-event. The frequency peaked during the event, and the topic discussion continued in the post event, thus with a slightly decreased frequency. Topics change over the span of the different windows. Therefore, there's a relation between the topics during the event and after the event, where the discussion topics continue. Different concepts have additional importance for a specific topic, based on the sentiments and emotions of the users. Further, based on the chosen timeframe of a three-week window for each of the events, the results have shown that the users tend to adopt a topic and join in on trending discussions rather fast. Consequently, the importance of a topic gains or loses importance and frequency over the windows. The word frequency analysis gave clear indication of recurring topics, which indicates that a majority of the tweets are related to the specific topics of the event. The researchers in Linton et al. (2017) found that online communities that are focused on cryptocurrencies are more likely to have strong word-of-mouth. In addition, they are more likely to discuss and form opinions on relevant events.

The third key result is related to the influence of Elon Musk. The findings in the Bitcoin and Dogecoin event analysis show that there is a window where Elon Musk's tweets have an impact, which is between the event and the post-event window. For example, in the Bitcoin event study, we see that the discussion regarding Elon Musk during the event is labeled 48% positive. In the post-event, the discussion amongst the users is still at 48% positive, but with a higher frequency of negatively charged words. The discussion regarding Elon Musk is 57% positive during the Dogecoin event. The most labeled emotion in this discussion is happy, followed by surprise. It is a temporary effect as the sentiment decrease to 53% positive in the post event. There is a window in time where the tweets from Elon Musk have a bigger impact, which is in during the event window. In the Dogecoin post event, the conversation topic changed from the main topic during the event to Musk's tweet about McDonald's. This resulted in "McDonalds" having a high frequency in the discussion. Moreover, this shows that Musk's tweets indeed have had an effect on the topics and discussions amongst the users, where a single tweet causes a shift in the discussion. Not on the greatest importance regarding the sentiments and emotions, but in terms of shifting the discussions and topics across the windows.

A study by Ante (2021) revelated that Elon Musk has a powerful influence on the cryptocurrency market and users. Due to the increasing number of people and the flow of information, it has become easier for a single user to influence and affect an entire market (Ante, 2021). For instance, the "Musk effect" phenomenon occurs when Elon Musk tweets about cryptocurrencies such as Bitcoin and Dogecoin and causes a price spike in the assets (Ante, 2021). It can be explained by the sudden increase in the price of cryptocurrencies after Musk's tweets about various topics, for instance, Bitcoin and Doge. When Musk tweeted that Tesla would accept Dogecoin as payment, Dogecoin spiked and traded up 23%. Musk's prolific use of social media and the increasing interest from amateur investors drove the price of Dogecoin higher. Since the currencies do not pay dividends, their returns are dependent on the increasing prices. If a well-known individual or company influences retail investors to buy cryptocurrencies, this could cause them to end up paying the highest price.

8 Limitations

The results of the research should be considered in the context of limitations. Aside from addressing the research questions, the process also generates additional aspects that can be explored in future research. One of the limitations of the research is the data collection. The scraped data only represents a small portion of the complete data. For instance, the Snscrape library can't pull all of the details related to the cryptocurrency interactions because the data is scraped based on keywords. Therefore, only interaction related to the keywords was extracted, even though the user interactions may go beyond this.

Another limitation is the timeframe of the data. The scraped tweets are between the time period of January 1st, 2020 till January 31st, 2022, due to the limited time on the thesis. Furthermore, because of the time constraints, a sentiment classifier was not built. Instead, lexicon-based sentiment analysis generated by VADER and TextBlob was utilized. Therefore, the results of the research may be affected by the lack of a sentiment classifier, because lexicon-based approaches can obtain lower accuracy than supervised techniques.

Only data from one social media platform was used. The limitation of the data source is that Twitter data often lack lengthy comments. For this reason, it can be useful to investigate other social media platforms that also have a high frequency of posts or social media platforms with high-quality posts at a lower frequency. In this way, the scope of the research can be investigated in more depth. In addition, the research was limited to four cryptocurrencies. Additional research could analyse more cryptocurrencies to compare user interactions.

The possible effects of Twitter bots on interactions are not researched due to the scope of the research. In addition, a heuristic approach was conducted. For this reason, the actual number of Twitter bots may be higher than the observed number. For instance, if a tweet refers to a giveaway, it is highly possible that it's tweeted by a bot. However, the tweet is only classified as a bot account if additional criteria is met. Therefore, the number of bots in the dataset may be higher than presented.

9 Conclusion and implications

The research aimed to study cryptocurrencies through the lenses of Twitter by applying various analytical methods to answer the research questions. The study used 1724328 tweets to research the user interaction on the cryptocurrencies Bitcoin, Ethereum, Dogecoin, and Uniswap. In order to answer the research questions, the Social set analysis framework was utilized, supported by sentiment and emotion analysis for further analysis on users and events. The analytical objective for conducting a Social set analysis was to identify user' interactions on Twitter with reference to cryptocurrencies.

The core findings show that user interactions indeed differ between larger and smaller coins, and that half of the sentiments were positive across all cryptocurrency datasets. However, smaller coins have more positive sentiments than larger coins, and mostly expressed happy emotions, while the larger coins mostly expressed fear. Further, the user analysis shows that there are more unique users in common between the larger coins, and that the users have a higher tweet frequency and volume compared to the unique users in common in the smaller coins. The findings in the event study revealed that the users have a temporal change in sentiments and emotions in the face of events and have significant impact on user interactions.

The findings have implications of practice for future studies on cryptocurrencies. For instance, extending the time frame for collecting data can allow researchers to conduct more comprehensive studies. In addition, the findings have implications of practice for small investors, interested parties, and those who are new to the world of cryptocurrency. The dynamic relationship between social media and user interaction on cryptocurrencies can influence decision-making. Social media is indeed a platform where users can share opinions, sentiments, and emotions despite the various points of view. Since the findings revealed that the cryptocurrencies have a close relationship to their respective categories, it can be valuable for investors to assess the different cryptocurrencies in context to each other.

Another implication of practice is that the analysis can help investors be more aware of their financial interests by studying the public's opinions about cryptocurrencies. Moreover, how the public's opinions, sentiments, and emotions differ between smaller coins and larger coins. On the other hand, we get to see the influence and effect of social media. Furthermore, the findings also have implications for social media analytics in terms of user-generated content and user-perspective in several fields of research. For instance, businesses may consider the various perspective of their users by extracting collective interactions and intelligence from Twitter users, and further utilizing the information to analyze their various characteristics.

The thesis contributes to demonstrating the possibilities and effectiveness of the Social set analysis framework to analyze and visualize a massive amount of social media data and user-generated data that is created in social media platforms such as Twitter. Another contribution is to the body of knowledge by explaining the connection between social media and cryptocurrency. In addition, the thesis contributes to literature on social media analytics by extracting, analyzing and interpreting social media data.

10 Bibliography

Abraham, Jethin, Daniel Higdon, John Nelson, and Juan Ibarra. “Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis.” *SMU Data Science Review* 1, no. 3 (2018): 1.

Ahmed, Wasim, Peter A. Bath, and Gianluca Demartini. “Using Twitter as a Data Source: An Overview of Ethical, Legal, and Methodological Challenges.” *The Ethics of Online Research*, 2017.

Al-Khafaji, Dr Hussein K., and Areej Tarief Habeeb. “Efficient Algorithms for Preprocessing and Stemming of Tweets in a Sentiment Analysis System.” *Journal of Computer Engineering* 19, no. 3 (2017): 44–50.

Alqaryouti, Omar, Nur Siyam, Zainab Alkashri, and Khaled Shaalan. “Cryptocurrency Usage Impact on Perceived Benefits and Users’ Behaviour.” In *European, Mediterranean, and Middle Eastern Conference on Information Systems*, 123–36. Springer, 2019.

Angeris, Guillermo, and Tarun Chitra. “Improved Price Oracles: Constant Function Market Makers.” In *Proceedings of the 2nd ACM Conference on Advances in Financial Technologies*, 80–91, 2020.

Ante, Lennart. “How Elon Musk’s Twitter Activity Moves Cryptocurrency Markets.” *Available at SSRN 3778844*, 2021.

Arifiyanti, Amalia Anjani, and Eka Dyar Wahyuni. “Emoji and Emoticon in Tweet Sentiment Classification.” In *2020 6th Information Technology International Seminar (ITIS)*, 145–50, 2020. https://doi.org/10.1109/ITIS50118.2020.9320988.

Aslam, Naila, Furqan Rustam, Ernesto Lee, Patrick Bernard Washington, and Imran Ashraf. “Sentiment Analysis and Emotion Detection on Cryptocurrency Related Tweets Using Ensemble LSTM-GRU Model.” *IEEE Access* 10 (2022): 39313–24. https://doi.org/10.1109/ACCESS.2022.3165621.

Asur, Sitaram, and Bernardo A. Huberman. “Predicting the Future with Social Media.” In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 1:492–99. IEEE, 2010.

Bartov, Eli, Lucile Faurel, and Partha Mohanram. “Can Twitter Help Predict Firm-Level Earnings and Stock Returns?,” 2018, 61.

Blau, Benjamin M., Todd G. Griffith, and Ryan J. Whitby. “Inflation and Bitcoin: A Descriptive Time-Series Analysis.” *Economics Letters* 203 (June 1, 2021): 109848. https://doi.org/10.1016/j.econlet.2021.109848.

Bohr, Jeremiah, and Masooda Bashir. “Who Uses Bitcoin? An Exploration of the Bitcoin Community.” In *2014 Twelfth Annual International Conference on Privacy, Security and Trust*, 94–101. IEEE, 2014.

Boukhers, Zeyd, Azeddine Bouabdallah, Matthias Lohr, and Jan Jürjens. “Ensemble and Multimodal Approach for Forecasting Cryptocurrency Price.” *ArXiv:2202.08967 [Cs, q-Fin]*, February 12, 2022. http://arxiv.org/abs/2202.08967.

Bowman, Robert G. “Understanding and Conducting Event Studies.” *Journal of Business Finance & Accounting* 10, no. 4 (1983): 561–84.

Bryman, Alan. *Social Research Methods*. 4th ed. Oxford ; New York: Oxford University Press, 2012.

Burggraf, Tobias, Toan Huynh, Markus Rudolf, and Mei Wang. "Do FEARS Drive Bitcoin?" *Review of Behavioral Finance* ahead-of-print (May 20, 2020). https://doi.org/10.1108/RBF-11-2019-0161.

Chatterjee, Ankush, Umang Gupta, Manoj Kumar Chinnakotla, Radhakrishnan Srikanth, Michel Galley, and Puneet Agrawal. "Understanding Emotions in Text Using Deep Learning and Big Data." *Computers in Human Behavior* 93 (2019): 309–17.

Conway, Luke. "Ten Important Cryptocurrencies Other Than Bitcoin." Investopedia, November 19, 2021. https://www.investopedia.com/tech/most-important-cryptocurrencies-other-than-bitcoin/.

De Vaus, D. A. *Research Design in Social Research*. London ; Thousand Oaks, Calif: SAGE, 2001.

DeMatteo, Megan. "Ethereum Continues to Hover Around $3,000. See How That Compares to Its Price History." *Time*, January 21, 2022. https://time.com/nextadvisor/investing/cryptocurrency/ethereum-price-history/.

Derbentsev, Vasily, Natalia Datsenko, Olga Stepanenko, and Vitaly Bezkorovainyi. "Forecasting Cryptocurrency Prices Time Series Using Machine Learning Approach." In *SHS Web of Conferences*, 65:02001. EDP Sciences, 2019.

Desai, Rashi. "How to Scrape Millions of Tweets Using Snscrape." *DataSeries* (blog), January 2, 2022. https://medium.com/dataseries/how-to-scrape-millions-of-tweets-using-snscrape-195ee3594721.

Dollarhide, Maya. "Social Media: Sharing Ideas and Thoughts." Investopedia, 2021. https://www.investopedia.com/terms/s/social-media.asp.

Domingo, Ribeiro-Soriano, Juan Piñeiro-Chousa, and M. Ángeles López-Cabarcos. "What Factors Drive Returns on Initial Coin Offerings?" Technological Forecasting and Social Change 153 (2020): 119915.

Detrixhe, John. "Everything You Need to Know about DeFi." Quartz, November 10, 2021. https://qz.com/2065446/everything-you-need-to-know-about-decentralized-finance-defi/.

Eaton, Tina. "DeFi Coins and Tokens: What Every Investor Should Know," 2021. https://www.kubera.com/blog/defi-coins.

Ekman, Paul, Robert W. Levenson, and Wallace V. Friesen. "Autonomic Nervous System Activity Distinguishes among Emotions." *Science* 221, no. 4616 (1983): 1208–10.

Elmer, Greg, Ganaele Langlois, and Joanna Redden. *Compromised Data: From Social Media to Big Data*. Bloomsbury Publishing USA, 2015.

Flesch, Benjamin, Ravi Vatrapu, Raghava Rao Mukkamala, and Abid Hussain. "Social Set Visualizer: A Set Theoretical Approach to Big Social Data Analytics of Real-World Events." In *2015 IEEE International Conference on Big Data (Big Data)*, 2418–27, 2015. https://doi.org/10.1109/BigData.2015.7364036.

Frankenfield, Jake. “Altcoin Investing: What Investors Need to Know.” Investopedia, 2021. https://www.investopedia.com/terms/a/altcoin.asp.

Frankenfield, Jake. “What Is Cryptocurrency?” Investopedia, 2022. https://www.investopedia.com/terms/c/cryptocurrency.asp.

Ganis, Matthew, and Avinash Kohirkar. *Social Media Analytics: Techniques and Insights for Extracting Business Value out of Social Media*. IBM Press, 2015.

Gao, Bingyu, Haoyu Wang, Pengcheng Xia, Siwei Wu, Yajin Zhou, Xiapu Luo, and Gareth Tyson. “Tracking Counterfeit Cryptocurrency End-to-End.” *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 4, no. 3 (November 30, 2020): 1–28. https://doi.org/10.1145/3428335.

Ghani, Norjihan Abdul, Suraya Hamid, Ibrahim Abaker Targio Hashem, and Ejaz Ahmed. “Social Media Big Data Analytics: A Survey.” *Computers in Human Behavior* 101 (December 1, 2019): 417–28. https://doi.org/10.1016/j.chb.2018.08.039.

Grover, Purva, Arpan Kumar Kar, Marijn Janssen, and P. Vigneswara Ilavarasan. “Perceived Usefulness, Ease of Use and User Acceptance of Blockchain Technology for Digital Transactions–Insights from User-Generated Content on Twitter.” *Enterprise Information Systems* 13, no. 6 (2019): 771–800.

Memes.” Investopedia, 2021. https://www.investopedia.com/wallstreetbets-slang-and-memes-5111311.

Hayes, Adam. “Blockchain Explained.” Investopedia, 2022. https://www.investopedia.com/terms/b/blockchain.asp.

Hayes, Adam. “What Is an Event Study?” Investopedia, 2022. https://www.investopedia.com/terms/e/eventstudy.asp.

Houben, Dr Robby, and Alexander Snyers. “Cryptocurrencies and Blockchain,” n.d., 103.

Hutto, Clayton, and Eric Gilbert. “Vader: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text.” In *Proceedings of the International AAAI Conference on Web and Social Media*, 8:216–25, 2014.

Johnson, Alison Grace. “What’s a Twitter Bot and How to Spot One,” 2020. https://us.norton.com/internetsecurity-emerging-threats-what-are-twitter-bots-and-how-to-spot-them.html.

Jussila, Jari, Eman Alkhammash, Norah Saleh Alghamdi, Prashanth Madhala, and Mohammad Ayoub Khan. “A Netnographic-Based Semantic Analysis of Tweet Contents for Stress Management.” *Computers, Materials and Continua* 70, no. 1 (2021): 1845–56.

Karalevicius, Vytautas, Niels Degrande, and Jochen De Weerdt. “Using Sentiment Analysis to Predict Interday Bitcoin Price Movements.” *The Journal of Risk Finance*, 2018.

Kendall, Will. “A Look Back in Time: Bitcoin Price History and Events Timeline | CoinMarketCap.” CoinMarketCap Alexandria, 2022. https://coinmarketcap.com/alexandria/article/bitcoin-price-history-and-events-timeline.

Kim, Evgeny, and Roman Klinger. “A Survey on Sentiment and Emotion Analysis for Computational Literary Studies.” *ArXiv Preprint ArXiv:1808.03137*, 2018.

Kim, Jooho, and Makarand Hastak. “Social Network Analysis: Characteristics of Online Social Networks after a Disaster.” *International Journal of Information Management* 38, no. 1 (2018): 86–96.

Kraaijeveld, Olivier, and Johannes De Smedt. “The Predictive Power of Public Twitter Sentiment for Forecasting Cryptocurrency Prices.” *Journal of International Financial Markets, Institutions and Money* 65 (2020): 101188.

Kumar, Anoop S., and Suvvari Anandarao. “Volatility Spillover in Crypto-Currency Markets: Some Evidences from GARCH and Wavelet Analysis.” *Physica A: Statistical Mechanics and Its Applications* 524 (2019): 448–58.

Lamon, Connor, Eric Nielsen, and Eric Redondo. “Cryptocurrency Price Prediction Using News and Social Media Sentiment.” *SMU Data Sci. Rev* 1, no. 3 (2017): 1–22.

Leech, Ollie. “What Is Uniswap? A Complete Beginner’s Guide - CoinDesk,” April 2, 2021. https://www.coindesk.com/business/2021/02/04/what-is-uniswap-a-complete-beginners-guide/.

Lex, Alexander, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister. “UpSet: Visualization of Intersecting Sets.” *IEEE Transactions on Visualization and Computer Graphics* 20, no. 12 (December 2014): 1983–92. https://doi.org/10.1109/TVCG.2014.2346248.

Linton, Marco, Ernie Gin Swee Teo, Elisabeth Bommes, C. Y. Chen, and Wolfgang Karl Härdle. “Dynamic Topic Modelling for Cryptocurrency Community Forums.” In *Applied Quantitative Finance*, 355–72. Springer, 2017.

Liu, Bing. “Sentiment Analysis and Opinion Mining,” 2012, 168.

Liu, Bing. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Second edition. Studies in Natural Language Processing. Cambridge ; New York: Cambridge University Press, 2020.

Locke, Taylor. “‘People Have Been Participating without Understanding the Risks’: Here’s What to Know about Cryptocurrency-Based DeFi.” CNBC, June 18, 2021. https://www.cnbc.com/2021/06/18/whats-defi-crypto-based-decentralized-finance-explained.html.

Loginova, Ekaterina, Wai Kit Tsang, Guus van Heijningen, Louis-Philippe Kerkhove, and Dries F. Benoit. “Forecasting Directional Bitcoin Price Returns Using Aspect-Based Sentiment Analysis on Online Text Data.” *Machine Learning*, November 18, 2021. https://doi.org/10.1007/s10994-021-06095-3.

Lomas, Natasha. “Duo Security Researchers’ Twitter ‘Bot or Not’ Study Unearths Crypto Botnet.” *TechCrunch* (blog), 2018. https://social.techcrunch.com/2018/08/06/duo-security-researchers-twitter-bot-or-not-study-unearths-crypto-botnet/.

Mahmoudi, Nader, Łukasz P. Olech, and Paul Docherty. "A Comprehensive Study of Domain-Specific Emoji Meanings in Sentiment Classification." *Computational Management Science*, August 18, 2021. https://doi.org/10.1007/s10287-021-00407-7.

Mai, Feng, Zhe Shan, Qing Bai, Xin Wang, and Roger HL Chiang. "How Does Social Media Impact Bitcoin Value? A Test of the Silent Majority Hypothesis." Journal of Management Information Systems 35, no. 1 (2018): 19–52.

Majumder, Prateek. "Analysing the Cryptocurrency of May 2021 in Python." *Analytics Vidhya* (blog), 2021. https://www.analyticsvidhya.com/blog/2021/05/analyzing-the-cryptocurrency-of-may-2021-python-for-finance-basics/.

Matilda, S. "Big Data in Social Media Environment: A Business Perspective." In *Decision Management: Concepts, Methodologies, Tools, and Applications*, 1876–99. IGI Global, 2017.

Menon, Karan, Hannu Karkkainen, Jari Jussila, Jukka Huhtamäki, Raghava Rao Mukkamala, Lester Allan Lasrado, Ravi Vatrapu, and Abid Hussain. "Analysing the Role of Crowdfunding in Entrepreneurial Ecosystems: A Social Media Event Study of Two Competing Product Launches." *International Journal of Entrepreneurship and Small Business* 33, no. 4 (2018): 575–606.

Mitra, Mallika. "Investing Has a Whole New Language — Here's Your Cheat Sheet." Money, 2021. https://money.com/investing-crypto-internet-language-glossary/.

Miraz, Mahdi H., and Maaruf Ali. "Applications of Blockchain Technology beyond Cryptocurrency." *ArXiv Preprint ArXiv:1801.03528*, 2018.

Mnif, Emna, Isabelle Lacombe, and Anis Jarboui. "Users' Perception toward Bitcoin Green with Big Data Analytics." *Society and Business Review* 16, no. 4 (January 1, 2021): 592–615. https://doi.org/10.1108/SBR-02-2021-0016.

Mukkamala, Raghava Rao, Jannie Iskou Sørensen, Abid Hussain, and Ravi Vatrapu. "Social Set Analysis of Corporate Social Media Crises on Facebook." In 2015 IEEE 19th International Enterprise Distributed Object Computing Conference, 112–21, 2015. https://doi.org/10.1109/EDOC.2015.25.

Naimy, Viviane, Omar Haddad, Gema Fernández-Avilés, and Rim El Khoury. "The Predictive Capacity of GARCH-Type Models in Measuring the Volatility of Crypto and World Currencies." *PloS One* 16, no. 1 (2021): e0245904.

Narman, Husnu S., and Alymbek Damir Uulu. "Impacts of Positive and Negative Comments of Social Media Users to Cryptocurrency." In *2020 International Conference on Computing, Networking and Communications (ICNC)*, 187–92. Big Island, HI, USA: IEEE, 2020. https://doi.org/10.1109/ICNC47757.2020.9049693.

Nothman, Joel. "Upsetplot Documentation," 2021.

O'Connor, Brendan, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series." In *Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

Oates, Briony J. *Researching Information Systems and Computing*. London ; Thousand Oaks, Calif: SAGE Publications, 2006.

Ortu, Marco, Stefano Vacca, Giuseppe Destefanis, and Claudio Conversano. "Cryptocurrency Ecosystems and Social Media Environments: An Empirical Analysis through Hawkes' Models and Natural Language Processing." *Machine Learning with Applications* 7 (2022): 100229.

Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010.

Park, Han Woo, and L. E. E. Youngjoo. "How Are Twitter Activities Related to Top Cryptocurrencies' Performance? Evidence from Social Media Network and Sentiment Analysis." *Drustvena Istrazivanja* 28, no. 3 (2019).

Phillips, Ross C., and Denise Gorse. "Predicting Cryptocurrency Price Bubbles Using Social Media Data and Epidemic Modelling." In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1–7. IEEE, 2017.

Piccolboni, Luca, Giuseppe Di Guglielmo, Luca P. Carloni, and Simha Sethumadhavan. "CRYLOGGER: Detecting Crypto Misuses Dynamically." In *2021 IEEE Symposium on Security and Privacy (SP)*, 1972–89, 2021. https://doi.org/10.1109/SP40001.2021.00010.

Piñeiro-Chousa, Juan, M. Ángeles López-Cabarcos, Aleksandar Sevic, and Isaac González-López. "A Preliminary Assessment of the Performance of DeFi Cryptocurrencies in Relation to Other Financial Assets, Volatility, and User-Generated Content." *Technological Forecasting and Social Change* 181 (August 1, 2022): 121740. https://doi.org/10.1016/j.techfore.2022.121740.

Poyser, Obryan. "Exploring the Determinants of Bitcoin's Price: An Application of Bayesian Structural Time Series." *ArXiv Preprint ArXiv:1706.01437*, 2017.

Rodeck, David. "An Introduction to Dogecoin, The Meme Cryptocurrency." Forbes Advisor, April 20, 2021. https://www.forbes.com/advisor/investing/what-is-dogecoin/.

Rousseau, Denise M., Joshua Manning, David Denyer, Copyright ©denise M. Rousseau, Joshua Manning, David Denyer, Denise M. Rousseau, et al. "Evidence in Management and Organizational Science: Assembling the Field's Full Weight of Scientific Knowledge Through Syntheses," 2008.

Sarkar, Dipanjan. *Text Analytics with Python: A Practitioner's Guide to Natural Language Processing*. Springer, 2019.

Sarkar, Tushar, and Nishant Rajadhyaksha. "TLA: Twitter Linguistic Analysis." *ArXiv Preprint ArXiv:2107.09710*, 2021.

Shahid, Mohammad. "Understanding the Shiba Inu Ecosystem | Cryptopolitan," October 14, 2021. https://www.cryptopolitan.com/shiba-inu-ecosystem-explained/.

Shahul, ES. "Sentiment Analysis in Python: TextBlob vs Vader Sentiment vs Flair vs Building It From Scratch." neptune.ai, 2021. https://neptune.ai/blog/sentiment-analysis-python-textblob-vs-vader-vs-flair.

Sharma, Radhika, Vandana Ahuja, and Shirin Alavi. "The Future Scope of Netnography and Social Network Analysis in the Field of Marketing." *Journal of Internet Commerce* 17, no. 1 (January 2, 2018): 26–45. https://doi.org/10.1080/15332861.2017.1423533.

Sharma, Rakesh. “Decentralized Finance (DeFi) Definition and Use Cases.” Investopedia, March 24, 2021. https://www.investopedia.com/decentralized-finance-defi-5113835.

Sloan, Luke, and Anabel Quan-Haase, eds. *The SAGE Handbook of Social Media Rsearch Methods*. Los Angeles ; London: SAGE reference, 2017.

Sovbetov, Yhlas. “Factors Influencing Cryptocurrency Prices: Evidence from Bitcoin, Ethereum, Dash, Litcoin, and Monero.” *Journal of Economics and Financial Analysis* 2, no. 2 (2018): 1–27.

Sponder, Marshall, and Gohar F. Khan. *Digital Analytics for Marketing*. Routledge, 2018.

Stieglitz, Stefan, Milad Mirbabaie, Björn Ross, and Christoph Neuberger. “Social Media Analytics–Challenges in Topic Discovery, Data Collection, and Data Preparation.” *International Journal of Information Management* 39 (2018): 156–68.

Tandon, Chahat, Sanjana Revankar, and Sidharth Singh Parihar. “How Can We Predict the Impact of the Social Media Messages on the Value of Cryptocurrency? Insights from Big Data Analytics.” *International Journal of Information Management Data Insights* 1, no. 2 (2021): 100035.

Thomas, Rhys. “What Is a Meme Coin?” The Face, March 11, 2021. https://theface.com/life/what-is-a-meme-coin-dogecoin-floki-shiba-inu.

Vatrapu, Ravi, Abid Hussain, Niels Buus Lassen, Raghava Rao Mukkamala, Benjamin Flesch, and Rene Madsen. “Social Set Analysis: Four Demonstrative Case Studies.” In *Proceedings of the 2015 International Conference on Social Media & Society*, 1–9, 2015.

Vatrapu, Ravi, Raghava Rao Mukkamala, Abid Hussain, and Benjamin Flesch. “Social Set Analysis: A Set Theoretical Approach to Big Data Analytics.” *Ieee Access* 4 (2016): 2542–71.

Vornewald, Kilian, Andreas Eckhardt, and Julia Krönung. “Emotions in Information Systems Research – A Five Component View,” n.d., 16.

Wallaker, Matthew. “The 7 Key Factors Influencing Cryptocurrency Value.” MUO, December 12, 2021. https://www.makeuseof.com/factors-influencing-the-cryptocurrency-value/.

Wang, Charlie, and Ben Luo. “Predicting $GME Stock Price Movement Using Sentiment from Reddit r/Wallstreetbets.” In *Proceedings of the Third Workshop on Financial Technology and Natural Language Processing*, 22–30. Online: -, 2021. https://aclanthology.org/2021.finnlp-1.4.

White-Gomez, Alex. “What Is Market Cap in Crypto?” Accessed May 13, 2022. https://www.one37pm.com/nft/what-is-market-cap-in-crypto.

Yin, Robert K. *Case Study Research: Design and Methods*. 4th ed. Applied Social Research Methods, v. 5. Los Angeles, Calif: Sage Publications, 2009.

Zhao, Dejin, and Mary Beth Rosson. “How and Why People Twitter: The Role That Micro-Blogging Plays in Informal Communication at Work.” In *Proceedings of the ACM 2009 International Conference on Supporting Group Work*, 243–52, 2009.

Zimmerman, Chris, Mari-Klara Stein, Daniel Hardt, Christian Danielsen, and Ravi Vatrapu. “EmotionVis: Designing a Tool for Emotion Text Inference and Visual Analytics,” 2016.

Uniswap Protocol. “Introducing UNI,” September 16, 2020. https://uniswap.org/blog/uni.

TWINT - Twitter Intelligence Tool. Python. 2017. Reprint, TWINT Project, 2021. https://github.com/twintproject/twint.

CoinMarketCap. “Cryptocurrency Prices, Charts And Market Capitalizations,” 2022. https://coinmarketcap.com/.

TradingView. “ETHUSD History — Timeline of Major Events,” 2022. https://www.tradingview.com/symbols/ETHUSD/history-timeline/.

“Tutorial: Quickstart — TextBlob 0.16.0 Documentation.” Accessed May 21, 2022. https://textblob.readthedocs.io/en/dev/quickstart.html#sentiment-analysis.

“Tweepy Documentation — Tweepy 4.10.0 Documentation.” Accessed May 21, 2022. https://docs.tweepy.org/en/stable/.

ethereum.org. “Decentralized Finance (DeFi).” Accessed November 24, 2021. https://ethereum.org.

ethereum.org. “What Is Ethereum?” Accessed November 24, 2021. https://ethereum.org.

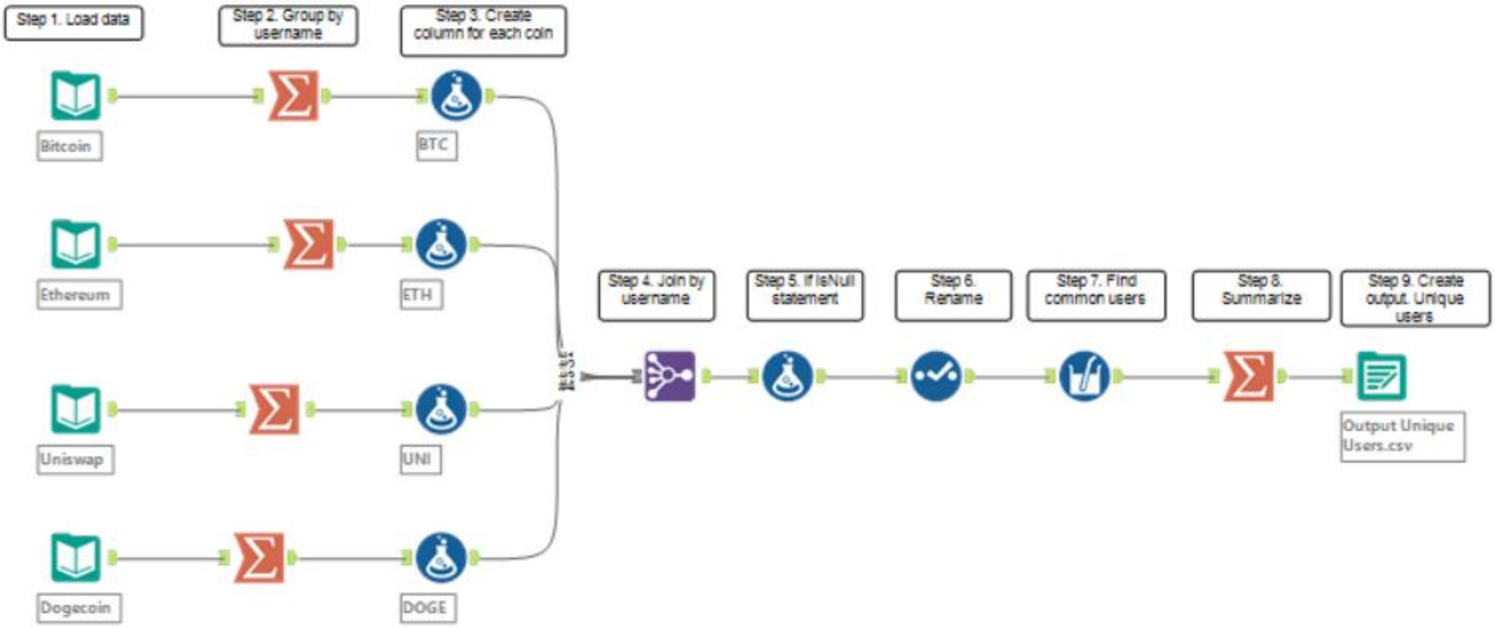
11 Appendices

Table of Contents

A. List of figures

B. List of Tables

C. Alteryx workflow



D. Code Implementation

1. Data scraping

```
#import libraries
import snscrape.modules.twitter as
sntwitter import pandas as pd

maxTweets = 20000

# List to append tweet data
tweets_list = []

keyword=['ethereum', 'eth']

for i,tweet in enumerate(sntwitter.TwitterSearchScraper(f'{keyword} + since:2020-01-01
until:2020-01-31 lang:en').get_items()):
    if i>maxTweets:
        break
    tweets_list.append([tweet.date, tweet.id, tweet.content, tweet.user.username])

# Creat dataframe from tweets list
tweets_df = pd.DataFrame(tweets_list, columns=['Datetime', 'Tweet Id', 'Text', 'Username'])
```

2. Data pre-processing

```
#import libraries
import preprocessor as p
import pandas as pd
import re
from nltk.tokenize import RegexpTokenizer

EMO_UNICODE = {}
UNICODE_EMO = {v: k for k, v in EMO_UNICODE.items()}

def convert_emojis(text):
    for emot in UNICODE_EMO:

text=re.sub(r'('+emot+')',"_".join(UNICODE_EMO[emot].replace(",","").replace(":","").split()
), text)
    return text
```

```
#Create space between emoji
import emoji

def extract_emojis(text):
   return ''.join((' '+c+' ') if c in emoji.UNICODE_EMOJI['en'] else c for c in text)

df['Text'] = df.Text.apply(extract_emojis)

# Passing both functions to 'Text'
df['Text'] = df.Text.apply(convert_emojis)

#Remove urls
def remove_urls(text):
   url_pattern = re.compile(r'https?://\S+|www\.\S+')
   return url_pattern.sub(r'', text)

df['Text'] = df['Text'].str.replace(r'\s*@\w+', '', regex=True)

def clean_data(dataframe):
#replace URL of a text
   df['Text'] = df['Text'].str.replace('http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[!*\(\),]|(?:%[0-9a-
fA-F][0-9a-fA-F]))+', ' ')

clean_data(df)
print(df['Text']);

#Remove stopwords
from gensim.parsing.preprocessing import remove_stopwords

def stopword_removal(row):
   Text = row['Text']
   Text = remove_stopwords(Text)
   return Text

df['Text'] = df.apply(stopword_removal, axis=1)

#Remove extra white spaces
df['Text'] = df['Text'].str.replace('\s\s+', ' ')

#Remove punctuation except apostrophe and exclamation mark
df["Text"] = df['Text'].str.replace('[^\w\s\!\’]','')

#Removes all numbers
```

```
df['Text'] = df['Text'].str.replace('\d+', '')
#Remove duplicates
df.drop_duplicates(subset ="Text", keep="first", inplace = True)

df = df.sort_values(by="Datetime")
df.head()
```

3. UpSet plot

```
#import libraries
import pandas as pd
from matplotlib import pyplot as plt
from upsetplot import plot

Users = pd.read_csv('Users.csv', sep=";")

grouping = Users.groupby(['ETH','BTC','US','DOG'])['Count_Users'].sum()

fig = plt.figure(figsize=(15, 7))
plot(grouping, fig=fig, show_counts=True, element_size=None)
plt.show()
```

4. Sentiment analysis

VADER

```
#import libraries
import nltk
nltk.download('vader_lexicon')
from nltk.sentiment.vader import SentimentIntensityAnalyzer
sid = SentimentIntensityAnalyzer()
nltk.download('punkt')
nltk.download('averaged_perceptron_tagger')
nltk.download('wordnet')
nltk.download('vader_lexicon')
from nltk.tokenize import TweetTokenizer
import pandas as pd

df= pd.read_csv('Ethereum-tweets.csv')
```

```
df['Text'] = df['Text'].astype('str')

df['scores'] = df['Text'].apply(lambda Text: sid.polarity_scores(Text))

df['compound'] = df ['scores'].apply(lambda score_dict: score_dict['compound'])
df ['sentiment_type']=''
df.loc[df.compound>0,'sentiment_type']='Positive'
df.loc[df.compound==0,'sentiment_type']='Neutral'
df.loc[df.compound<0,'sentiment_type']='Negative'
df.head()
```

Textblob

```
#import libraries
import pandas as pd
import numpy as np
import textblob as TextBlob # for doing sentimental analysis
import re # regex for cleaning the tweets
from nltk.sentiment.vader import SentimentIntensityAnalyzer
from nltk.sentiment.util import
from textblob import TextBlob
from nltk import tokenize


df= pd.read_csv('Ethereum-tweets.csv')
from textblob import TextBlob
df["sentiment_score"] = df["Text"].apply(lambda x: TextBlob(str(x)).sentiment.polarity)
df["sentiment"] = np.select([df["sentiment_score"] < 0, df["sentiment_score"] == 0, df["sentiment_score"] > 0],
                      ['Negative', 'Neutral','Positive'])
```

5. Emotion analysis

```
#import libraries
import pandas as pd
import text2emotion as te
import seaborn as sns

df= pd.read_csv('Ethereum-tweets.csv')

emotion_2= []
for i in df["Text"].values.tolist():
  emotions = te.get_emotion(i)
  # print(emotions)
  keymax =max(emotion, key=emotion.get)
  emotion_2.append(keymax)
#print(emotion_2)

df['Text2emotion']= emotion_2


df['Text2emotion'].value_counts()
df['Text2emotion'].describe()


#plot
sns.set_theme(style="whitegrid")
sns.countplot(df['Text2emotion'],order
= df["Text2emotion"].value_counts(normalize=True).index)
```

6. User analysis

```
#Tweet volume
df['Datetime'] = pd.to_datetime(df['Datetime'])
vol = df.groupby(pd.Grouper(key='Datetime',freq='d')).size().reset_index(name='tweet_vol')
vol['tweet_vol'].describe()

#Volume plot
```

```
vol['Datetime'] = pd.to_datetime(vol['Datetime'])
vol = vol.set_index('Datetime')
plt = plt.subplot2grid((5,4), (0, 0), rowspan=3, colspan=4)
plt.plot(vol.index, vol["tweet_vol"])
plt.gcf().set_size_inches(10,10)

#Prices
import warnings
warnings.filterwarnings('ignore')  # Hide warnings
import datetime as dt
import pandas as pd
pd.core.common.is_list_like = pd.api.types.is_list_like
import pandas_datareader.data as web
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import matplotlib.dates as mdates

import plotly.express as px
start = dt.datetime(2020, 1, 1)
end = dt.datetime(2022,1,31)

#UNI
uni = web.DataReader("UNI1-USD", 'yahoo', start, end)
uni.reset_index(inplace=True)
crypto= uni[['Date','Adj Close']]
crypto= crypto.rename(columns = {'Adj Close':'UNI'})
crypto["UNI"]= uni["Adj Close"]

# ploting the adjusted closing price
fig, axs =plt.subplots(2,2,figsize=(16,10),gridspec_kw ={'hspace': 0.2, 'wspace': 0.1})
axs[0,0].plot(crypto[UNI])
axs[0,0].set_title(UNI)
```

```
axs[0,1].plot(crypto['DOGE'])
axs[0,1].set_title('DOGE')
plt.show()
```

7. Event study

```
df['Text2emotion'].value_counts()
df['Text2emotion'].describe()
df['sentiment'].value_counts()
df['sentiment'].describe()
df['Username'].describe()
df['Username'].value_counts()

from matplotlib import pyplot as plt
from upsetplot import plot
fig = plt.figure(figsize=(15, 7))
plot(grouping, fig=fig, show_counts=True, element_size=None,facecolor="darkblue")
plt.show()
```

8. Descriptive analysis

```
#import libraries
import pandas as pd
import nltk
from nltk.tokenize import word_tokenize
from nltk.tokenize import RegexpTokenizer
from nltk.stem import WordNetLemmatizer
from nltk.corpus import stopwords
from gensim.parsing.preprocessing import remove_stopwords
from collections import Counter
from nltk import ngrams
from collections import Counter
import matplotlib.pyplot as plt
import seaborn as sns
```

```
df= pd.read_csv('Ethereum-tweets.csv')
df['Text'] = df['Text'].str.lower()
df['Text'].describe()
print(stopwords.words('english'))

def stopword_removal(row):
    Text = row['Text']
    Text = remove_stopwords(Text)
    return Text

df['Text'] = df.apply(stopword_removal, axis=1)

df['tokens'] = df.apply(lambda row: nltk.word_tokenize(row['Text']), axis=1)
df.head()

# Generate the Bitcoin N-grams where N=2
df_Text = ' '.join(df.Text)
df_processed = word_tokenize(df_Text)
df_ngrams = Counter(ngrams(df_processed, n=2))
print(dict(df_ngrams.most_common(10)))

# Generate top 10 words
def token_count(tokens, N=10):
    """Returns the top N tokens from the frequency count"""
    return Counter(tokens).most_common(N)

# Get the top 10 words
df_top10 = token_count(df_processed)
df_top10

#Visualization histogram
df= df['Text']
df = pd.DataFrame(df)
description_list = df['Text'].values.tolist()
word_frequency = Counter(" ".join(description_list).split()).most_common(10)
```

```
# Most frequent token returns a list of (word, count) tuples
words = [word for word, _ in word_frequency]
counts = [counts for _, counts in word_frequency]

plt.figure(figsize=(15,8))
plt.bar(words, counts, color= 'blue')
plt.title("10 most frequent tokens Ethereum")
plt.ylabel("Frequency")
plt.xlabel("Words")
plt.show()

#calculating average tweet length and word count
import numpy as np
df['text_len'] = df['Text'].astype(str).apply(len)
df['text_word_count'] = df['Text'].apply(lambda x: len(str(x).split()))
print("Average length of tweets ", round(np.mean(df['text_len'])))
print("Average word counts of tweets", round(np.mean(df['text_word_count'])))

df['text_word_count'].describe()
df['text_len'].describe()

%matplotlib inline
plt.figure(figsize=(5,5))
doc_lens = [len(d) for d in df.Text]
plt.hist(doc_lens, bins = 100)
plt.xlim([0, 300]);
plt.ylabel('Number of Tweets')
plt.xlabel('Tweets character length')
sns.despine();

# time series
sns.set(style='whitegrid', palette='muted', font_scale=1.2)
reactions = data.groupby(['Datetime']).count()
ax = reactions.Text.plot(figsize=(10,5),ls='-',color='blue')
ax.xaxis.grid(True)
ax.yaxis.grid(True)
```